First System Architecture Description

Project acronym:	**SENSE4US**
Project full title:	**Data Insights for Policy Makers and Citizens**
Grant agreement no.:	**611242**
Responsible:	**Juri Papay**
Contributors:	**Steve Taylor, Paul Walland**
Document Reference:	**D3.2**
Dissemination Level:	**PU**
Version:	**FINAL**
Date:	**06/07/15**

Project co-funded by the European Union under the Seventh Framework Programme

History

Version	Date	Modification reason	Modified by
0.1	30/03/2015	Initial draft	J.Papay
0.2	08/04/2015	Architecture development process	J.Papay
0.3	21/04/2015	Scenarios	J.Papay
0.4	07/05/2015	Mapping scenarios into architecture components	J.Papay
0.5	19/05/2015	Architecture of the prototype, non-functional requirements	J.Papay
0.6	08/06/2015	1st complete draft	Steve Taylor
0.8	15/06/2015	Updated complete draft	Steve Taylor
1.0	19/06/15	Submitted for review	Steve Taylor
1.0TW	29/06/16	Review Comments	Timo Wandhöfer
1.1	30/06/15	Address Review Comments	Steve Taylor
1.3	06/07/15	Final Review & Address Comments	Paul Walland, Steve Taylor

Table of contents

List of figures

List of tables

List of abbreviations

Abbreviation	Description
DoW	Description of Work
WP	Work package

Executive summary

In this deliverable we have presented the architecture of the Sense4Us system. In addition, we have presented a process to determine the architecture that takes into account a set of differing inputs:

- functional requirements derived from end user requirements;
- the original concept of the project and the project's initial architectural principles; and
- the initial demonstration prototype.

The architecture has been derived based on D3.1 (Functional Specification), which itself is in turn based on D2.1 (user requirements). The Description of Work (DoW) is also a key input to the process as it determines the project's overall concept. The project's concept includes the project's research work conducted by partners in WPs 4, 5 and 6, and so this research was also taken into account. The architecture also has background work as input in the form of a basic infrastructure that has been proved in previous project and has been developed further within Sense4us, to produce D7.1, the initial demonstrator.

The architecture itself has been determined through the definition of use case scenarios that represent the core functionality of the project and how it may be used by end users. Functional requirements were analysed to remove redundancy and to determine the ones that needed addressing, and requirements were allocated to the scenarios that could address them. For each scenarios, a "sub-architecture" was created, which shows the components from the initial demonstrator and the components from the research workpackages. Each scenario's sub-architecture shows how these components may be used to deliver the functionality described by the DOW. This process highlighted where any components were missing, and these have been described. The final step in the architectural determination process was to combine all the scenario sub-architectures into one master architecture, which is the outcome of this deliverable. This is the first version of the system architecture and will be updated in month 30.

1 Introduction

This document describes the first design of the Sense4us system. It also describes the process by which the design was arrived at. As an input for this deliverable we have used the functional specification from D3.1, the user requirements from D2.1, the original concept of the project from the Description of Work (DoW) and the prototype of Sense4us system. The aim of this effort was to ensure that the requirements and objectives of the project were fully addressed and the architecture could deliver the required functionality.

The current document is the first version of the system architecture and will be updated in month 30 when more details about the individual tools such as Sentiment Analysis, LOD search and Policy Modelling become available. These tools are currently under development therefore they are described only on the input/output and functionality levels.

The next section sets out some background definitions for the term “architecture”, and after this, the major structure of the document follows:

- Section 2 describes the process of deriving the Sense4us system – how the architecture fits into the overall project structure
- Section 3 introduces the key principles of architecture design - these form the basis of the architecture
- Section 4 describes the determination of the Sense4us architecture, putting the process described in section 2 into action.
- Section 5 concludes with a summary.

1.1 Background and Definitions

In the literature we can find numerous definitions for the term *“Software architecture”*, but we will use the following definition.

> Software architecture encompasses the set of significant decisions about the organization of a software system:
>
> a) Selection of the structural elements and their interfaces by which a system is composed
> b) Behaviour as specified in collaborations among those elements
> c) Composition of these structural and behavioural elements into larger subsystems
> d) Architectural style that guides this organization
>
> Software architecture also involves functionality, usability, resilience, performance, reuse, comprehensibility, economic and technology constraints, trade-offs and aesthetic concerns[1].

[1] Per Kroll and Philippe Kruchten. The Rational Unified Process Made Easy: A Practitioners Guide to the RUP, Addison Wesley 2003, page 315.

The software architecture serves various purposes[2]:

- a blueprint for both the system and the project developing it
- defines the work assignments for the design and implementation teams
- specifies system qualities such as performance, modifiability, and security, none of which can be achieved without a unifying architectural vision
- a vehicle for early analysis to make sure that the design approach will yield an acceptable system
- it is the conceptual glue that holds every phase of the project together.

In this document we also focus on capturing the rationale why certain decisions were taken and what the implications are of these decisions. For the Sense4us architecture design we have followed the recommendations of ISO standard *"Systems and software engineering — Architecture description (ISO/IEC/IEEE 42010 First edition 2011-12-01)"*[3]. The key recommendations of this standard can be summarised as follows[4]:

1) The architecture as such is defined as: "**architecture** (system) - fundamental concepts or properties of a system in its environment embodied in its elements, relationships, and in the principles of its design and evolution"
2) "An architecture description includes one or more architecture views. An architecture view (or simply, *view*) addresses one or more of the concerns held by the system's stakeholders."

The standard provides a long list of possible concerns such as: functionality, feasibility, usage, system purposes, system features, system properties, known limitations, structure, behaviour, performance, resource utilization, reliability, security, information assurance, complexity, evolvability, openness, concurrency, autonomy, cost, schedule, quality of service, flexibility, agility, modifiability, modularity, control, inter-process communication, deadlock, state change, subsystem integration, data accessibility, privacy, compliance to regulation, assurance, business goals and strategies, customer experience, maintainability, affordability and disposability."

Naturally we are not going to address all these concerns and provide a detailed design for each of them, but focus on aspects such as usage, behaviour, system features and properties.

Although the standard provides definitions for the key concepts, it makes no recommendations on how the architecture itself should be determined. We find that this is a shortcoming of this standard and a more detailed specification for the determination of software architectures would be desirable. As a result of this, we have determined that we

---

[2] David Garlan. Software Architecture: a Roadmap. In The Future of Software Engineering, A. Finkelstein,ed., ACM Press, 2000.

[3] Systems and software engineering — Architecture description, ISO/IEC/IEEE 42010, First edition 2011-12-01, http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6129467

[4] Systems and software engineering — Architecture description, ISO/IEC/IEEE 42010, First edition 2011-12-01, http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6129467, page 2.

need a process for the system architecture design for Sense4us, and this process is described in the next Section.

2 Process for Determination of the Sense4us Architecture

There are a number of processes to determine system architectures for industrial applications, for example as discussed in Hofmeister et al[5]. Sense4us has a slightly different context than for industrial software engineering as it is a research project and has a different rationale and starting points from more traditional software engineering projects. Therefore it has specific needs regarding a process to determine its system design. Hence our approach is directed at producing system designs for research projects, and we believe that this process may be applicable to other research projects.

We determined that the process for creating the Sense4us system design should not only utilise the functional requirements derived in D3.1 from the end users' requirements. Our assertion is that other aspects need to be considered as well, such as the original concept of the project and the research objectives of the academic partners. The main reason for this is that the majority of the requirements from end users were more concerned with how information was delivered, rather than the actual information itself. For example the top-scoring requirement is that information should contain as much provenance information as possible. This is perfectly reasonable, but it was felt that if the architecture addressed only the functional requirements, it is possible that the core concept of the project and its research would be overlooked.

We have therefore determined a process that addresses these objectives, and this process is described in this section (see Figure 1). This process also allows us to check the correctness of the design and to make sure that both the user and functional requirements are addressed.

The main classes of input for the process are:

1) In the architecture design we have utilised the information and knowledge that has been accumulated in the previous deliverables and also integrated the contributions from the project partners. In D2.1 the user requirements were identified. They provided input for deriving the functional requirements as described in D3.1, First Functional Requirements Description. D3.1 has also prioritised the functional requirements and identified the key architecture components.
2) In addition to the functional requirements, other aspects needs to be taken into account in the architecture design. The project's original concept as described in the DoW is one of the key inputs of the process. The research work performed by the academic partners in WPs 4, 5 and 6 also needs to be taken into account, as it represents the major novel components of the project. Also initial infrastructural design choices need to be taken in to account to determine if they are sound and can therefore be included in the final design.

The process itself is depicted by Figure 1, next. Figure 1 in essence describes the entire Sense4us project capturing the relationships and interactions between individual WPs. It

[5] Christine Hofmeister et al. "A general model of software architecture design derived from five industrial approaches". The Journal of Systems and Software 80 (2007) 106–126.

allows us to see the place of architecture development in context of the other WPs, and how the other WPs contribute to this activity.

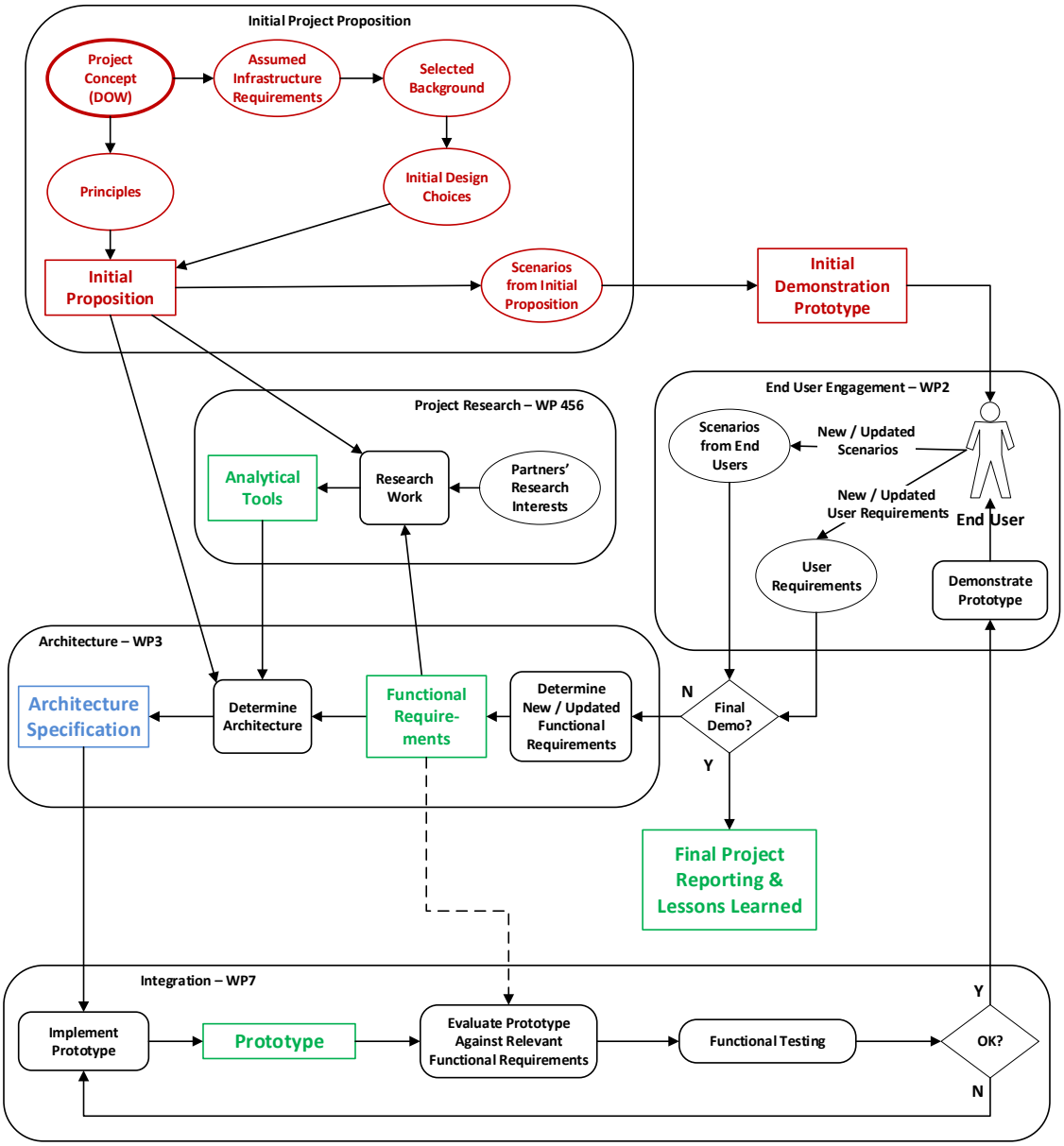


Figure 1: Process for Determination of the Sense4us System Design[6]

The starting point of the Initial Project Proposition effort is the "Project Concept" described in the DOW. The "Assumed Infrastructure Requirements" describe infrastructural requirements such as data storage, security and user interface that are needed to deliver the envisaged functionality. Based on these assumed requirements, the concept of the project and the previous background work of the project partners, we can select background components that contribute towards these requirements. These steps produce a set of "Initial Design Choices". These are combined with the overall principles of the project (its core concept as described at the proposal stage and represented in the DOW), to produce the Initial

[6] In Figure 1 the ellipses represent activities and the square boxes documents or software.

Proposition. This sets out the starting point for the system design in the project, as it contains the position before the project started. It is fully accepted that the Initial Proposition will inevitably be changed as the project progresses.

The output of the "Initial Project Proposition" effort is embodied in the "Initial Demonstration Prototype", which was demonstrated at the first review, and enhanced and delivered at PM18 (D7.1). This prototype is used for demonstrating the project's overall concept (a toolkit to help policy making researchers) and for engaging the end users (see WP2) in the requirement specification process.

The Initial Proposition also contains the initial approaches for the Research Work in WP4, WP5 and WP6 from the DOW, and provides an input to the first version of the "Architecture". The Research Work is about creating novel analytical tools that help the policy makers, and clearly the nature of these tools has a direct impact on the architecture. From the perspective of the architecture, the tools are treated as independent modules and the architecture should provide facilities for the integration, replacement or upgrade of these tools.

The purpose of "End User Engagement" is to evaluate the prototype and to provide / update end-user requirements. This is the work of WP2 and is where the initial prototype is tested, and feedback from end users provides new or updated end user requirements. WP2 also provides scenarios – situations and use cases where the user interacts with the system.

The "Architecture" effort (WP3 – this WP) provides the specification of the system architecture that serves as a blueprint for software development, coding and testing. The major task in this group is "Determine Architecture", and this takes into account the Initial Proposition, the Analytical Tools and the Functional Requirements that are derived by taking into account of the Scenarios and the User requirements. It should be also noted the Functional Requirements influence the Research Work done in the project, for example, where a previously-unconsidered requirement is determined important by the project and within the scope of the respective research partner's research agenda. The outcome of the "Determine Architecture" process is the actual Architecture Specification.

The objective of "Integration" (WP7) is to produce a prototype of the system from the Architecture Specification. This is evaluated against the functional requirements and tested whether it provides the expected functionality. If the outcome of testing is satisfactory the prototype is demonstrated to the end user otherwise the implementation cycle gets repeated.

There is an overall iteration cycle that is the lower half of the picture. This represents repeated Functional Requirements -> Architecture Specification -> Implementation -> Testing -> Demonstration cycles. These will be iterated as many times as time and resources allow within the project, so that the system is as fully specified and consistent with user requirements as possible. The exit point of the cycle is the "Final Demo". This is the final demonstration of the toolkit created by the project, and the final evaluation report of the project will make recommendations for exploitation beyond the end of the project.

2.1 Process for Architecture Determination

This section concerns what happens in the "Determine Architecture" box in Figure 1. A detailed process diagram is shown in Figure 2. This is a snapshot example applicable at the time of writing (because it uses current deliverables), and the diagram is likely to be updated in the future.

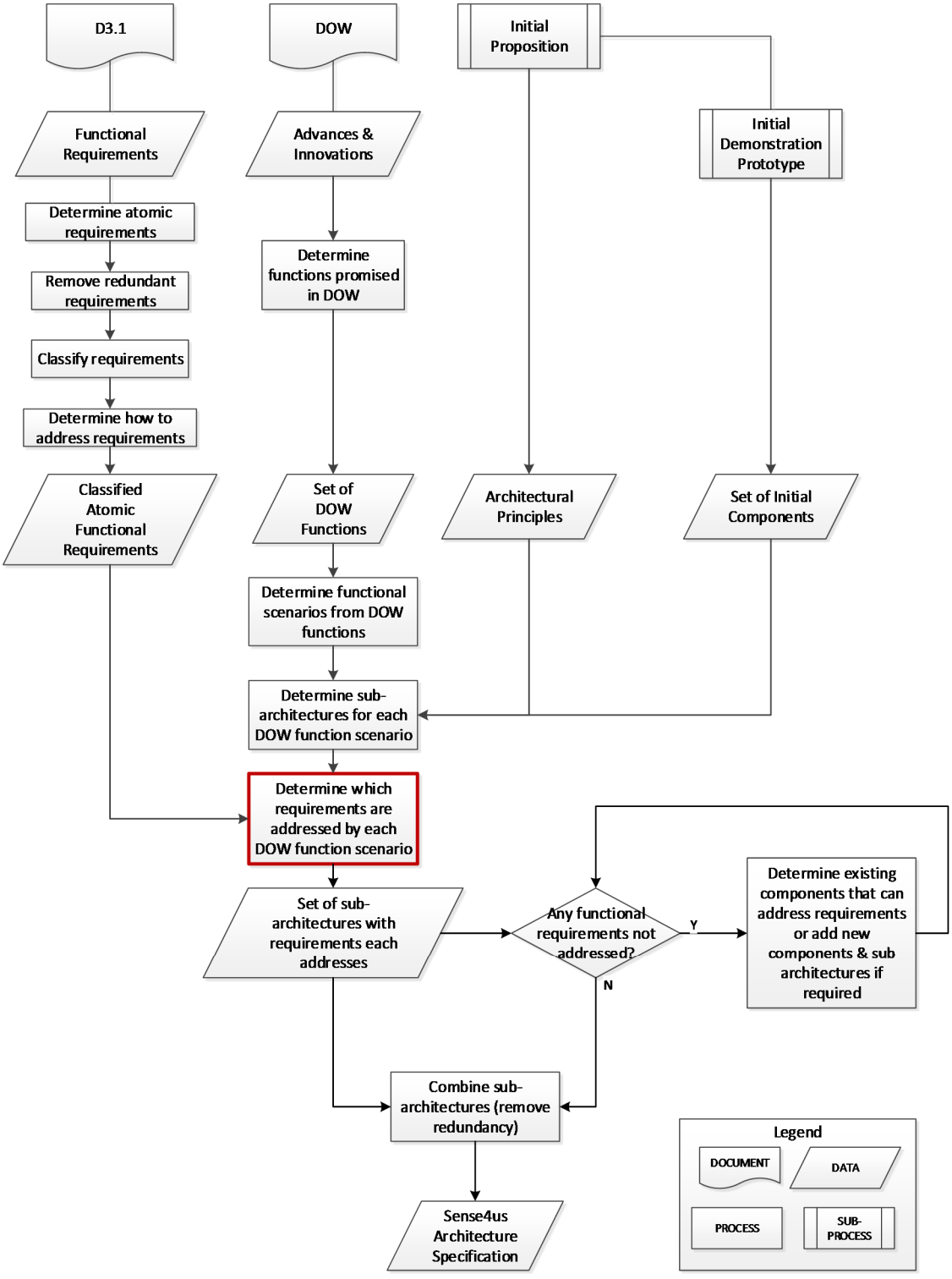


Figure 2: Process for Determination of Architecture

The process begins with three elements:

- D3.1, which contains functional requirements.
- The DOW, which contains the original concept of the project, and how it will advance the state of the art.
- The Initial Proposition, as defined above, which includes the architectural principles and the Initial Demonstration Prototype.

The functional requirements from D3.1 are obviously needed, as they represent the functionality required to address the end users' requirements. These are processed to remove any redundancy, and to determine how to address them. The result is a set of atomic requirements with notes on how they will be addressed.

The DOW contains advances and innovations, which are statements of how the project will advance the state of the art, and these also need to be represented in the architecture design as they represent the original concept of the project. These are analysed to determine the functionality needed to deliver the advances and innovations. The result of this is a set of so-called "DOW functions", which represent the concept of the project.

The Initial Proposition contains an initial architectural starting point, based on background infrastructure brought into the project (shown below in Figure 3). This needs to be taken into account, as it is the basis of the implementation. This has been developed further into the demonstrator described in D7.1, which contains a set of components, and all this needs to be taken into account when developing the system architecture. The Initial Proposition also contains core architectural principles (discussed in more detail in Section 3), which also need to be evaluated before inclusion into the system design.

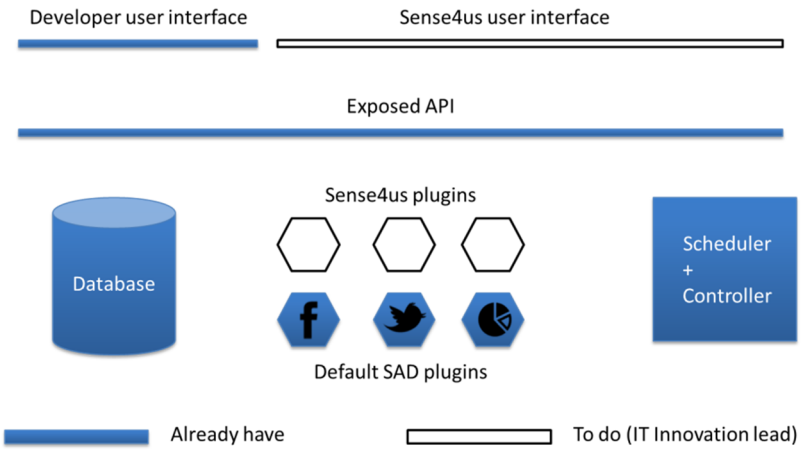


Figure 3: Initial infrastructure concept presented at the kick-off meeting

Hence, the architecture design is based on four elements:

- Classified atomic functional requirements (and how to address them)
- A set of DOW functions representing the innovations of the project
- Architectural principles
- A set of initial components

For each DOW function, a scenario is determined that illustrates how an end user can get the benefit described by the DOW function. These so-called "DOW function scenarios" are the means by which the four elements above are combined and evaluated. The essence of using scenarios derives from use cases, described as follows:

> "The most critical functionality of the system is captured in the form of scenarios (or use cases). By critical we mean: functions that are the most important, the raison d'être of the system, or that have the highest frequency of use, or that present some significant technical risk that must be mitigated."[7].

A "sub-architecture" is determined for each DOW function scenario that determines how the initial components can be combined to address the functionality needed. This shows how the existing components can be put together to address the needed functionality and also identifies any functionality that cannot be addressed by a component, which will necessitate the addition of new components or the extension of existing ones. The sub-architecture also follows the architectural principles and any highlights any deficiencies in them, which may mean they need to be modified. The red process box in the diagram highlights a major junction in the process. This is where the DOW functions are assessed to determine which functional requirements they address.

After this point, any functional requirements not addressed by any DOW function scenarios are identified. Each one of these unaddressed scenarios is evaluated to determine how it can be addressed, and whether the functionality it represents has an impact on the architecture, for example if new components are needed or an additional sub-architecture is needed. Once this is done, the sub-architectures are combined into one single architecture.

The next section described the architectural principles that are incorporated and tested within the architectural determination process, and Section 4 describes the process in action by determining the Sense4us architecture.

[7] Kruchten, Philippe (1995, November). Architectural Blueprints — The "4+1" View Model of Software Architecture.IEEE Software 12 (6), pp. 42-50.

3 Architecture Design Principles

In this section we describe the governing principles that were followed during the architecture design. These principles originate from the project concept itself and from the accumulated expertise of previous projects such as WeGov[8], Experimedia[9], BonFIRE[10] etc. These principles can be summarised as follows:

a) Data centricity
b) Extensibility
c) Web-hosted architectural pattern

In the following sub-sections we discuss the key principles of Sense4us architecture design. At the time of writing this report we have already developed a prototype of Sense4us system that is mainly used for explaining the concept of the project and for requirement gathering. This prototype implements only a fraction of the project's potential functionality, but it has implemented the following key principles of architecture design and can be considered as an embryonic version of the system.

3.1 Data centricity

Data centricity plays crucial role in the design. The essence if this concept is that data is at the core of the toolkit. Tools operate on data, produce transformed data and other tools can operate on this. All data types i.e. documents, metadata, provenance or other information are all stored in a database (see Figure 4). Users can view data items, and where compatible, users can choose which tool they want to operate on a piece of data. The arguments supporting the data centric approach are following:

a) Enables modular decomposition
b) Provides a basis for a flexible infrastructure
c) Easier integration of new tools (including third party tools if necessary)
d) Simplified interfaces and interactions between modules

[8] http://www.wegov-project.eu/

[9] http://www.experimedia.eu/

[10] http://www.bonfire-project.eu/

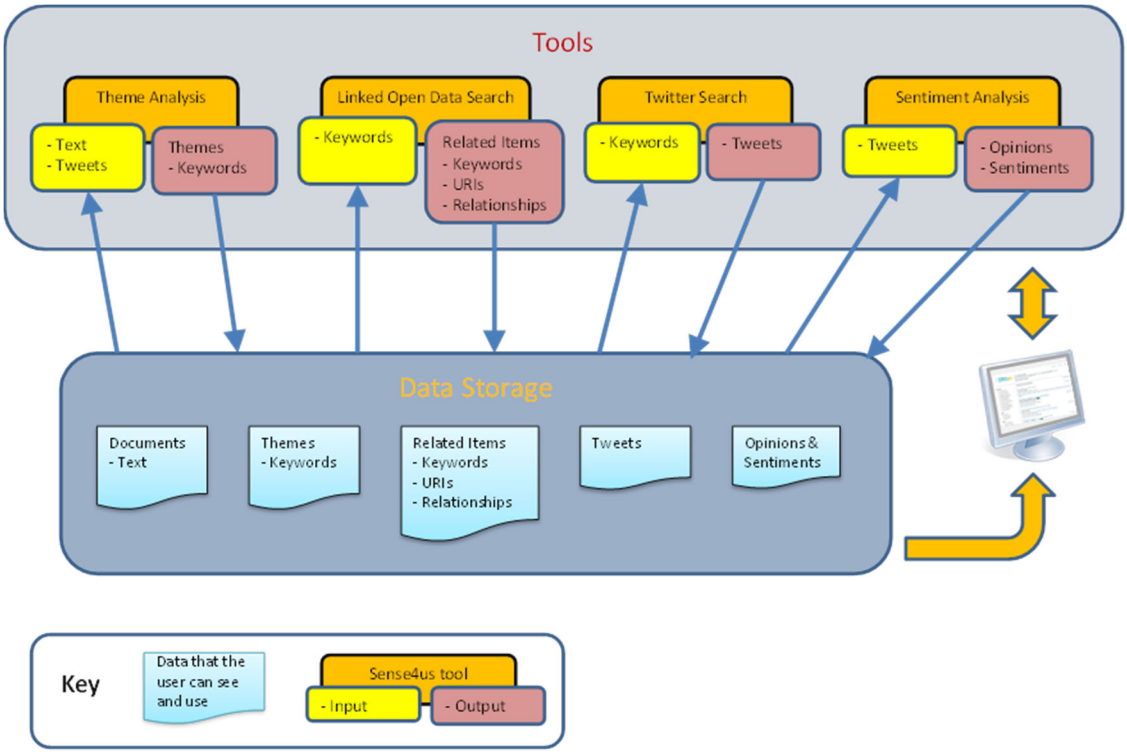


Figure 4: Data centric concept of Sense4us system

Figure 4 shows an example of the data centric concept with some (not all) Sense4us tools and data storage. The essence of the concept is that the user can run a tool can run on any compatible data at any time they wish. The tools can be used individually but they can be also linked up (via the data they consume and produce) into a sequence in order to realise more complex scenarios.

Access to data is restricted to the user that owns the data. This means that data uploaded to Sense4us by a user from outside or generated by a user running a tool belongs to that user, and no other user can see or alter that data. Users may share data with other users, and this mean that permitted users can read the data. If another user wants to use the data, they must copy it, as they must not alter data owned by someone else.

The data centric view is the overall philosophy for the architecture design, and a more detailed description is found in deliverable D7.1[11].

3.2 Extensibility

Extensibility represents the ability to plug-in new tools, replace or update them. The key questions in this context are *"How do we add new tools?"* and *"How to match up tools and data?"* In the context of Sense4us, tools are software artefacts that perform major functions, and most tools are the outputs of the research work in WPs 4, 5, and 6. Other notable tools include searches.

[11] D7.1 Sense4us - First Demonstration System, April 2015.

For integration of the tools, we have adopted an approach that uses wrappers that perform interface functions between the infrastructure and a tool:

- formats input data for the tool if needed
- keeps track of all metadata (job status, for example)
- runs the tool
- formats the output
- saves output into the database

The wrapper pattern means that we can use a standardised architecture, and for each new tool we write a new wrapper that does the actions specific to the tool being wrapped. Details of wrappers and how they are applied to the tools present in the Sense4us system at the current time can be found in the DOW function scenarios in Section 4.3.

For the integration of data and the tools in the Sense4us architecture we will utilise the data-centric approach, described above. The input of a tool always comes from the DB, and a tool's output is always stored in the DB as well. This means that one tool's output can be used as the input for another tool, when they are compatible. Tools' outputs can also be presented to the user via the UI (see Figure 5).

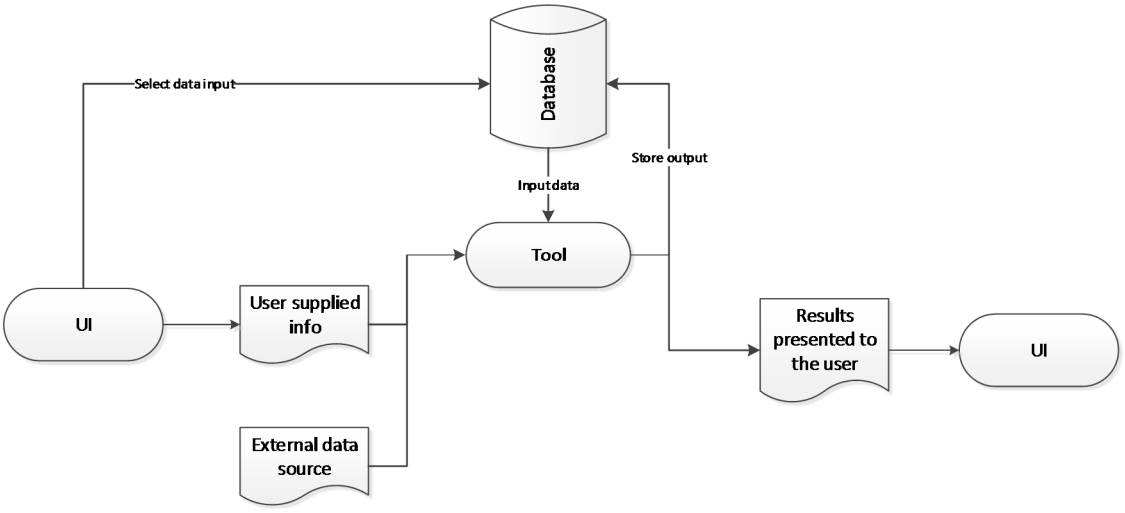


Figure 5 – Origin and destination of Sense4us data

To determine which tools can use which data, we propose to use the mechanism of *tagging*. The idea is that a tool needs to know what data it can work with and also be able to attach a tag (identifier) to the data that it produces. A tag in general describes the type of the data, for example a data can be a keyword, data graph etc. As a result of tagging we can build up a vocabulary of mappings that are used for matching the data and the tools that can use a particular data type (see Table 1).

Data tag	Tools that generate the output	Tools that can work with the given type of data	Comments
keyword	Theme Analysis, User typing in the UI	Social Media Search,	Generic keywords representing different aspects such as document

		LOD Search, Model Builder	themes, search terms etc.
overlap	Find overlaps		Shows the overlap between what people are saying on social media compared to what is said in a policy document. Enables the user to find out about the public's reaction to a policy.
sentiment	Sentiment Analysis	Model Builder	Extracting sentiment from a collection of tweets.
tweets	Social Media Search	Sentiment Analysis, Find overlaps	Records extracted from Twitter.
data graph	LOD Search	Model Builder	A data graph output from LOD search.
policy model	Policy Simulator		The policy model describes the factors and the correlations between them.

Table 1: Vocabulary of data tags

Each data type is tagged, and so the available actions for that item of data can be known, and this can be passed onto the user in the UI. For example, the user can click on a data item and can be presented a list of tools that can work with the given data.

3.3 Web-Hosted Service

At the time of writing, the project has already completed a demonstrator prototype of the Sense4us system (D7.1) that is mainly used for explaining the project's overall concept to end users and for requirement gathering. Its architecture is shown in Figure 6.

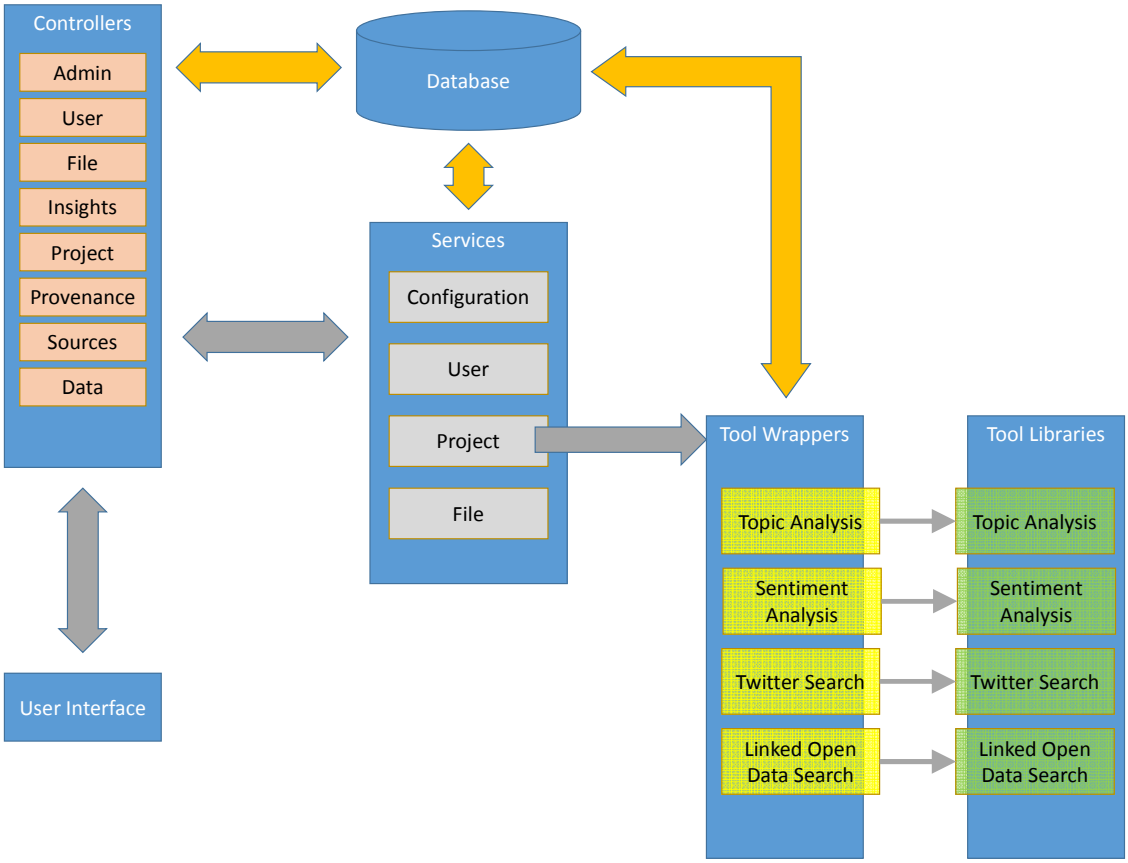


Figure 6: Sense4us prototype architecture

The core infrastructure is based on a hosted server that may be accessed using nothing more than a web browser via https. The primary aim for this is to avoid the need for users to install client software on their machines, which is seen as a barrier to adoption of the toolkit and is not always easy to do in public authorities or parliaments. This was derived from background work done by IT Innovation, and has been further developed in Sense4us. As such, it forms part of the "initial proposition" discussed in Section 2. The Framework provides a standardised pattern, involving the following main component types:

a) ***Database*** – this is for storing all data generated in the Sense4us system. This includes user data, e.g. from searches or analyses run by users, and management information, e.g. tracking information or tool specifications.
b) ***Controllers*** – these are function-specific components that receive/respond to requests from the UI.
c) ***Services*** – these are components that contain the bulk of the infrastructure's logic. They are used by controllers for functions such as fetching data and running tool wrappers, initialising the prototype on start-up etc.
d) ***Tool wrappers*** – these are components that enable different tools to be integrated. A wrapper behaves as the interface between the Sense4us infrastructure, in that it typically reads the submitted data (parameters, configuration etc), runs the tool and saves the output in the database.
e) ***Tool Libraries*** – these are the actual libraries that form the tools of Sense4us. These include Twitter Search, Topic Analysis, Sentiment Analysis, etc. The tools are often

outputs from the research WPs (4, 5 and 6), but may come from other sources, including third party libraries from outside the project.

f) ***Utility tools*** – these are ancillary components that perform specific functions such as determining location, language, handling file uploads, saving files in the DB, etc. These are not shown in Figure 6 for clarity.

g) ***User Interface*** – the web-application that the user interacts with.

The main controllers and services are discussed in more detail next.

3.3.1 Controllers

Controllers expose the backend API's endpoints to the UI. The controllers follow the REST standard[12], for example, a GET request to "/results" returns all results, POST request to "/results" creates a new result. The two main purposes for controllers are control of processing and management of data. The current list of controllers in the prototype includes:

a) ***AdminController*** – this performs administration functions: list users, delete a user, create new user, change user password, reset the service, or loading new users in bulk from a file.

b) ***DataController*** – this provides access to details of data in the database resulting from collections or analyses. It provides access to data items such as tweets, paragraphs, sentences, files, and analysis results. It provides a presentation of the data to the UI component so that the user can do things like: list items, get a single item, delete an item, tweet collection management, LOD results management or download a file.

c) ***FileUploadController*** – this is responsible for managing file uploads. It enables the user to specify a file to upload, and saves it into the database using the File Service.

d) ***InsightsController*** – this is responsible for accessing the results of analyses, for example main themes, sentiment, and theme overlaps.

e) ***ProjectsController*** – this is responsible for enabling the user to run different tools via the Project Service.

f) ***ProvenanceController*** – this is the presentation layer for provenance records that record the history of what processing the user has performed on data, and showing any provenance metadata about a data item.

g) ***SourcesController*** – this presents metadata about data sources. It provides access to data behind UI sources elements for data such as: tweets, files, and linked open data searches.

h) ***UserController*** – this is responsible for access to user related metadata and language setting.

[12] http://spring.io/understanding/REST

3.3.2 Services

"Services" are used by the "controllers". They provide various functionality, for example, matching users in the database to the submitted credentials. The present prototype contains these services:

a) ***ConfigurationService*** – this manages configuration of the whole Sense4us system for administrators including user management, prototype initialisation, DB configuration, shutdown etc.
b) ***FileService*** – this manages saving files along with relevant metadata in the DB, extracting paragraphs from files and saving them in the DB (for the purposes of summarisation by the Theme Analysis tool) and handling different file formats such as PDF, txt, or Word.
c) ***ProjectService*** – this is the main interface to tools, via tool-specific wrappers. The project service executes the tools in the wrappers and records provenance information in the DB for the instance of the wrapper execution (e.g. when it was executed, who did it, the parameters passed to the wrapper etc).
d) ***UserService*** – this is responsible for managing user credentials, and matching them against the data that the users own. It is used by all controllers to match credential submitted by the user's browser (i.e. their login name and password) against users in the database.

3.3.3 Wrappers

A Sense4us prototype wrapper's role is to interface between the standardised Sense4us infrastructure and specific tools. The wrappers main function is to execute tools given inputs from users and to record the output. Specific functionality of wrappers includes the following.

- A wrapper ensures correct input and parameters are submitted to the component. For example, it translates files or tweets into an array of strings instead for theme analysis.
- A wrapper ensures tools' output is saved in a format usable for presentation to the user and interpretable by other wrappers (so that one tool's output can be used as another's input).
- Some wrappers create additional output data if required, such as positive/negative tweets in addition to sentiment by sentiment analysis component for example.
- A wrapper stores the output data of a tool in the database. The wrapper also protects the ownership of the new data by assigning it to the user account of the user who ran the tool, meaning no one else can access the data (unless it is shared by the owner).
- A wrapper also stores a provenance record for any data output from the tool. This includes information such as: the tool's name, who executed the tool, the configuration and input data to the execution, the time and date etc.

4 Determination of Sense4us Architecture

This section describes the how the process described in Section 2.1 has been applied in the case of Sense4us to determine its architecture. Section 4.1 describes how functionality required from the DOW has been determined, Section 4.2 describes how the functional requirements from D3.1 have been analysed, and Section 4.3 describes DOW function scenarios with sub-architectures and requirements. Section 4.4 determines the functional requirements not address by and DOW function scenario, and how to address these requirements. Finally the overall architecture model is described in Section 4.5.

4.1 Functionality derived from initial project principles

In this section we list the scenarios that have been derived from the initial project principles. The primary source of functionality from the DOW is the section describing the advances made by the project, and the key innovations. For each advance (abbreviated to "DOW-ADV-") and innovation, we derive its key functionality, and this functionality is the basis of subsequent analysis. In the following analysis, the functionality derived is given an identifier number, prefixed with "DOW-FN-", standing for "DOW function".

4.1.1 The advances brought by the SENSE4US project

The following advances are described in the Sense4us DOW, Part B pages 5-6. They incorporate the essence of the project's original objectives.

DOW-ADV-1: Find and select available information relevant to the policy under development

Functions needed:

- DOW-FN-1: Intelligent Search of Heterogeneous Sources
- DOW-FN-2: Find Related Information

DOW-ADV-2: Link and homogenise the data to make it accessible and useful

Functions needed:

- DOW-FN-3: Combine Results from Heterogeneous Data Sources
- DOW-FN-4: Consistent Representation of Heterogeneous Data Sources

DOW-ADV-3: Model the policy against its objectives and intended impact

Functions needed:

- DOW-FN-5: Policy Model Construction
- DOW-FN-6: Simulation of Policy Option Impact

DOW-ADV-4: Validate the policy against existing legislation and broader over-arching policy

Functions needed:

- DOW-FN-7: Policy Validation in Context of Previous Similar Policy

DOW-ADV-5: Discover and take account of the views and opinions of non-governmental groups and the general public

Functions needed:

- DOW-FN-1: Intelligent Search of Heterogeneous Sources
- DOW-FN-8: Social Media Search
- DOW-FN-9: Social Media Discussion Analysis

DOW-ADV-6: Predict and test economic outcomes to ensure beneficial results

Functions needed:

- DOW-FN-6: Simulation of Policy Option Impact
- DOW-FN-10: Simulation of Economic Impact

DOW-ADV-7: Model and predict the likely social impact of policy

Functions needed:

- DOW-FN-5: Policy Model Construction
- DOW-FN-6: Simulation of Policy Option Impact
- DOW-FN-11: Simulation of Social Impact

DOW-ADV-8: Build a record of the policy development process in order to justify decisions made

Functions needed:

- DOW-FN-12: Recording of Information Discovery History
- DOW-FN-13: Provenance Information Extraction and Display
- DOW-FN-18: Recording of Policy Development History

DOW-ADV-9: Provide multiple policy options to be modelled, improving the negotiation or bargaining process between key stakeholders

Functions needed:

- DOW-FN-6: Simulation of Policy Option Impact
- DOW-FN-14: Evaluation of Multiple Policy Options

4.1.2 Innovations

4.1.2.1 Innovation 1:

“Intelligent search of all applicable and available governmental and external data sources, including linked open data, websites and social networks based on a topic-based analysis of a policy proposal used as a seed document”.

“Analysis of seed documents to determine search terms for a given policy document, including related areas”.

Source: Sense4us DoW Innovation 1, Part B page 37.

Functions needed:

- DOW-FN-1: Intelligent Search of Heterogeneous Sources
- DOW-FN-15: Generation of Search Terms
- DOW-FN-16: Extract Themes from Documents
- DOW-FN-2: Find Related Information

4.1.2.2 Innovation 2:

Analysis and extraction of discussion dynamics, sentiment and evidence from large data sets.

"Extraction of supporting evidence for the policy maker from vast amounts of semantic information available as linked data, open government data, domain specific sources and public policy discussion on social networking sites."

Source: Sense4us DoW Innovation 2, page 38.

Functions needed:

- DOW-FN-17: Evidence Extraction
- DOW-FN-1: Intelligent Search of Heterogeneous Sources
- DOW-FN-9: Social Media Discussion Analysis

4.1.2.3 Innovation 3:

"Modelling and simulation in order to predict the impact on society of a draft policy if it were enacted. Provision of a means of visualising the impact on society of policy decisions."

"Users can investigate the impact of different versions of the policy on society by using the simulator to construct "what-if" scenarios, taking into account different viewpoints and perspectives, multiple objectives and multiple stakeholders."

"Finding techniques for the validation of the simulation results."

Source: Sense4us, DoW, page 38.

Functions needed:

- DOW-FN-6: Simulation of Policy Option Impact
- DOW-FN-14: Evaluation of Multiple Policy Options

4.1.2.4 Innovation 4:

We are proposing a novel addition to the traditional policy making process, incorporating:

- Iterative development of a draft policy with multiple stakeholders
- More and better quality information and data from multiple sources than before ***(Covered in Innovation 1)***
- More advanced modelling and simulation than hitherto ***(Covered in Innovation 3)***

The benefits therefore are that the draft policy is better tested than before with multiple groups of stakeholders at different stages of development.

Functions needed:

- DOW-FN-13: Provenance Information Extraction and Display
- DOW-FN-18: Recording of Policy Development History
- DOW-FN-1: Intelligent Search of Heterogeneous Sources
- DOW-FN-2: Find Related Information
- DOW-FN-6: Simulation of Policy Option Impact
- DOW-FN-14: Evaluation of Multiple Policy Options

Sense4us, DoW, page 39.

4.1.2.5 Innovation 5:

Inclusion of public debate over social networks in policy process, supported by evidence from open data. Use of this outcome to support policy modelling or independently.

Functions needed:

- DOW-FN-17: Evidence Extraction
- DOW-FN-8: Social Media Search
- DOW-FN-9: Social Media Discussion Analysis
- DOW-FN-5: Policy Model Construction

4.1.2.6 Innovation 6:

Mashing of related but heterogeneous data from existing open governmental data, open linked data, historical information and discussion over Social Networks

Functions needed:

- DOW-FN-3: Combine Results from Heterogeneous Data Sources
- DOW-FN-4: Consistent Representation of Heterogeneous Data Sources

Source: Sense4us DoW Part B , page 39.

4.1.3 DOW Functions

The analysis in the above sections resulted in a set of functionality derived from the DOW, shown below.

- DOW-FN-1: Intelligent Search of Heterogeneous Sources
- DOW-FN-2: Find Related Information
- DOW-FN-3: Combine Results from Heterogeneous Data Sources
- DOW-FN-4: Consistent Representation of Heterogeneous Data Sources
- DOW-FN-5: Policy Model Construction
- DOW-FN-6: Simulation of Policy Option Impact
- DOW-FN-7: Policy Validation in Context of Previous Similar Policy
- DOW-FN-8: Social Media Search
- DOW-FN-9: Social Media Discussion Analysis
- DOW-FN-10: Simulation of Economic Impact
- DOW-FN-11: Simulation of Social Impact

- DOW-FN-12: Recording of Information Discovery History
- DOW-FN-13: Provenance Information Extraction and Display
- DOW-FN-14: Evaluation of Multiple Policy Options
- DOW-FN-15: Generation of Search Terms
- DOW-FN-16: Extract Themes from Documents
- DOW-FN-17: Evidence Extraction
- DOW-FN-18: Recording of Policy Development History

Looking at these functions, it was apparent that some contributed to others, so the functions were organised in a diagram to show their relationships, and this is shown in Figure 7.

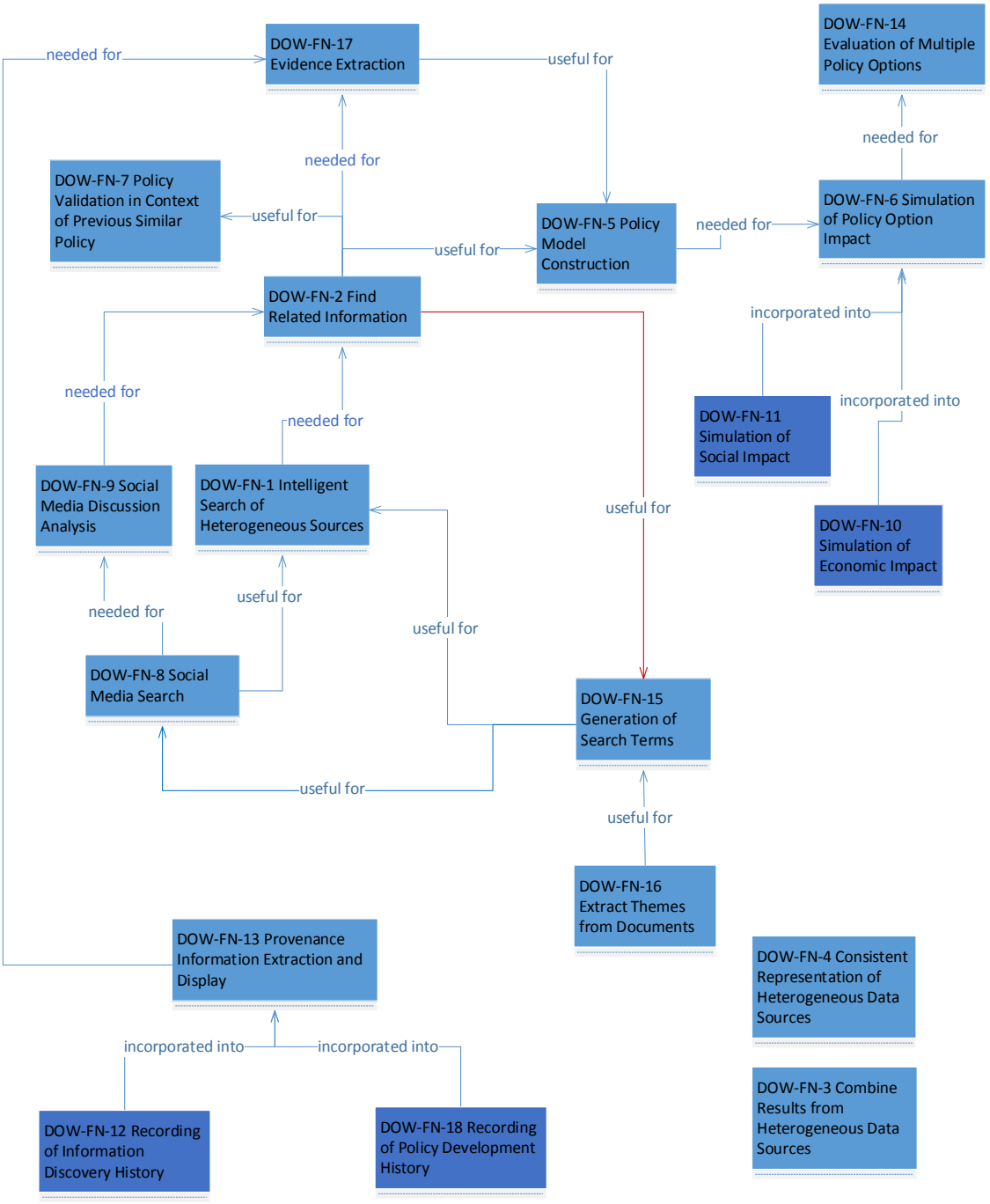


Figure 7: DOW Function Hierarchy

Figure 7 shows the DOW functions, and how they are related. A major relationship is the so-called "useful for" relationship. This means that one function is useful in assisting another function's behaviour. For example, DOW-FN-16, extract themes from documents, is useful for DOW-FN-15, generation of search terms, because DOW-FN-16 suggests search terms based on analysis of documents and finding keywords describing the key themes of the document. A key property of the "useful for" relationship is that is not mandatory – DOW-FN-15 will still be able to do its work if DOW-FN-16 does not contribute anything. In this case, the user is free to choose their own search terms in DOW-FN-15, and the terms supplied by the theme analysis

of DOW-FN-16 can be taken as search terms to be used, or suggestions to be ignored – it is up to the user to choose.

Working up the stack of "useful for" relationships, we can see that DOW-FN-15, generation of search terms, is useful for DOW-FN-1, Intelligent Search of Heterogeneous Sources, and DOW-FN-8, Social Media Search. This is clearly because DOW-FN-15 supplies search terms for both types of searches.

Looking at the relationship between DOW-FN-8, Social Media Search, and DOW-FN-9, Social Media Discussion Analysis, we see that DOW-FN-8 is "needed for" DOW-FN-9. This is a relationship where the dependent function cannot do its work unless the independent function supplies its output. Clearly in this case, the dependent function, social media discussion analysis, cannot work unless it has input, which needs to be generated by a social media search.

The search of heterogeneous sources in DOW-FN-1 comprises LOD search, and the social media search of DOW-FN-8 contributes to it. DOW-FN-1 is needed for DOW-FN-2, Find Related Information, because the search is locating information related to search terms related to the policy subject area (or for any area the user chooses – this is dependent on the search terms chosen by the user). DOW-FN-9, social media discussion analysis is also needed for DOW-FN-2 because the social media discussion analysis locates additional terms that people talk about when they mention the policy subject area.

Finding related information to policy subject areas is a major aspect of the toolkit, enabling the user to discover information not just within their policy subject area, but around it as well, for example what people also mention when taking about the subject area. This brings in other factors that the policy maker may not know about. Therefore DOW-FN-2 is an important function. By providing related factors to the policy subject area, and how they are related, DOW-FN-2 is needed for DOW-FN-17, Evidence extraction, DOW-FN-2 is also useful for the construction of policy models in DOW-FN-6.

Finding related information itself is useful for DOW-FN-7, validation of the policy in the context of previous similar policies. The user can select the current policy under development, together with a previous policy, perform similar analyses on both and compare the results.

There is a "feedback" arrow (coloured red in the diagram to highlight it) connecting DOW-FN-2 (find related information) and DOW-FN-15 (generation of search terms). This is to indicate that the functions may be used iteratively and at any time the user wants – the user can find related information around a policy subject area, and this may give them new ideas for search terms, which may be used in subsequent searches to find further information.

DOW-FN-13 is the recording and display of provenance information. It incorporates two other more specific DOW functions, DOW-FN-12 (recording of information discovery history) and DOW-FN-18 (recording of policy development history). These two functions are coloured with a darker blue box to indicate that they are incorporated into another function. Provenance information is important for evidence extraction, as provenance gives the user information to

enable them to make a judgement about whether they believe a piece of information or not, and so DOW-FN-13 is needed for DOW-FN-17, evidence extraction.

DOW-FN-5, policy model construction is needed for DOW-FN-6, simulation of policy option impact. It is necessary because a policy model is the subject of the simulation. DOW-FN-6, simulation, incorporates two other more specific simulation functions, DOW-FN-10 (simulation of economic impact) and DOW-FN-11 (simulation of social impact). DOW-FN-6 (simulation) is useful for the evaluation of multiple policy options (DOW-FN-14), because different policy options can be simulated and evaluated.

DOW-FN-17 (evidence extraction) is concerned with helping the user structure the information they have discovered. In this scenario, the user can build an information model using results from previous analyses, most notably DOW-FN-2 (find related information) and DOW-FN-13 (extract provenance information). Combination of results (DOW-FN-3) is partially addressed by allowing the user to choose which data they wish to use when creating their information models. It is for this reason DOW-FN-3 is incorporated into DOW-FN-17

The architectural pattern of wrappers that stores output data partially addresses DOW-FN-4 (consistent representation of data), in that each different type of output data can be stored in the database along with provenance information describing its origins and processing. This architectural pattern is incorporated in many of the DOW functions, for example DOW-FN-16 (theme analysis), DOW-FN-8 (social media search), DOW-FN-1 (search of heterogeneous data sources). DOW-FN-4 is not described separately, but its incorporation into these functions is not shown for reasons of clarity of the main diagram – instead it is shown in a separate figure, Figure 8 below.

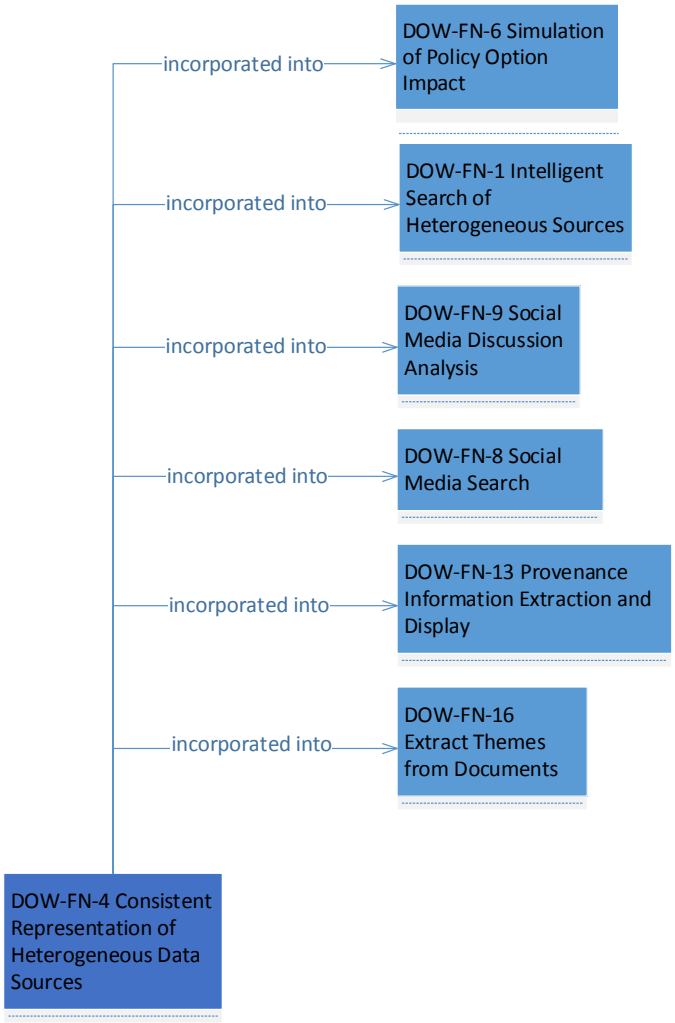


Figure 8: Incorporation of DOW-FN-4 into other DOW functions

4.2 Functional Requirements Analysis

This section contains the analysis of the functional requirements from D3.1. The broad process for the analysis took the following form.

- Firstly, some requirements from D3.1 actually contained more than one requirement, so these were decomposed into individual atomic requirements.
- There was some redundancy in the D3.1 requirements, so these were removed by identifying the requirements that duplicated functionality and deleting one of them (usually the one with the lowest priority[13] is deleted – if this is not the case, the remaining function assumes the higher of the priorities).
- Next there was determination of any requirements that will not be addressed, giving reasons.

[13] It should be noted that contrary to convention, here priorities are actually scores, so a larger number is of higher priority than a smaller number. This is because priorities are determined by collating the priorities (high, medium and low) the end user partners put on a requirement in D3.1.

- The requirements were organised into a structure, because it became apparent that there were groups of requirements, and within these groups, there was a hierarchy from general to specific requirements.
- From the hierarchy, a definitive list of requirements to be addressed was derived. For each of these requirements, a brief statement about how the project will address the requirement was made.

Each of these steps is discussed next, with associated results.

4.2.1 Determine Atomic Requirements

Firstly, the set of functional requirements from D3.1 was analysed to determine atomic requirements from them. The analysis process was simple: look for conjunction or disjunction words in the original requirement, and where they occurred, split the requirement into atomic separate requirements. The output of this process is shown in Table 2, next, which shows the original requirements and the resulting new atomic requirements. The requirements in Table 2 are the basis for all subsequent analysis in this document. Each atomic requirement retains its origin ID, so the atomic requirements can be traced back to their original. The convention for identifying atomic requirement ID is simply to suffix the original ID with “a”, “b” and “c” to denote each of the atomic requirements deriving from the original. Some requirements are not addressed. These are highlighted in red, and reasons for not addressing them are given in the discussion after the table.

Original Requirement			**Atomic Decomposition**		
F-ID	**Description**	**Pri-ority**	**A**	**B**	**C**
F-1	**Search different data formats and interface to multiple data sources**	6	F-1a - Search different data formats (Priority = 6)	F-1b - Search multiple data sources (Priority = 6)	
F-2	**Add new data sources**	6	F-2a - Add new data sources (Priority = 6)		
F-3	**Enable the user to add specific data sources and websites as search targets**	8	F-3a - Enable the user to add specific data sources (Priority = 8)	F-3b - Enable the user to add specific websites as search targets (Priority = 8)	
F-4	**Enable traditional online media (e.g. news sites) to be searched**	6	F-4a - Enable traditional online media (e.g. news sites) to be searched (Priority = 6)		
F-5	**Enable opinions from social media and other forums to be gathered**	7	F-5a - Enable opinions from social media to be gathered (Priority = 7)	F-5b - Enable opinions from other forums to be gathered (Priority = 7)	

Original Requirement			Atomic Decomposition		
F-ID	**Description**	**Pri-ority**	**A**	**B**	**C**
F-6	**Enable searching multiple data sources with the same query and mouse click from the user**	10	F-6a - Enable searching multiple data sources with the same query and mouse click from the user (Priority = 10)		
F-7	**Enable finding information for subjects related to those in the policy**	7	F-7a - Enable finding information for subjects related to those in the policy (Priority = 7)		
F-8	**Enable the user to filter search results - include or exclude results according to a wide range of criteria**	9	F-8a - Enable the user to filter search results (include or exclude results according to a wide range of criteria) (Priority = 9)		
F-9	**Present to the user all available information about the provenance of some data that has been found by the tool been found by the tool**	22	F-9a - Show all available provenance information for a piece of data (Priority = 22)	F-9b - (Show history of what user has done with data inside the toolkit) (Priority = 22)	
F-10	**Enable the user to restrict search results to those that they trust**	6	F-10a - Enable filtration of data (Priority = 6)	F-10b - Provide mechanism to help user make a trust judgement about data (Priority = 6)	F-10c - Enable user to indicate trusted data sources (Priority = 6)
F-11	**Enable the user to sort data by different criteria**	3	F-11a - Enable the user to sort data by different criteria (Priority = 3)		
F-12	**Easy to understand documentation and training on what the tools are capable of, how to use the tools, and the types of result to expect**	11	F-12a - Documentation (Priority = 11)	F-12b - Training (Priority = 11)	
F-13	**Transparency in the creation and operation of the policy models**	6	F-13a - Transparency in the creation of the policy models (Priority = 6)	F-13b - Transparency in the operation of the policy models (Priority = 6)	
F-14	**Enable the searching for scientific information relevant to the policy subjects**	3	F-14a - Enable the searching for scientific information relevant to the policy subjects (Priority = 3)		

Original Requirement			Atomic Decomposition		
F-ID	**Description**	**Pri-ority**	**A**	**B**	**C**
F-15	**Enable the collection of sentiments / opinions about the policy subjects**	3	F-15a - Enable the collection of sentiments about the policy subjects (Priority = 3)	F-15b - Enable the collection of opinions about the policy subjects (Priority = 3)	
F-16	**The policy maker should be able to examine a policy model and to add new factors if necessary**	6	F-16a - The policy maker should be able to examine a policy model (Priority = 6)	F-16b - The policy maker should be able to add new factors to policy model (Priority = 6)	
F-17	**Enable searching for factors related to their policy subjects using Sense4Us from different data sources**	9	F-17a - Enable searching for factors related to their policy subjects (Priority = 9)	F-17b - Enable searching for factors from different data sources (Priority = 9)	
F-18	**Acquire information from social media on the topics and themes of the policy in question**	4	F-18a - Acquire information from social media on the topics and themes of the policy in question (Priority = 4)		
F-19	**Discover any publicly available characteristic information about the social media participants who are expressing the opinions**	5	F-19a - Discover any publicly available characteristic information about the social media participants who are expressing the opinions (Priority = 5)		
F-20	**Enable the user to specify different time ranges for when the data was (or is) relevant in a search**	3	F-20a - Enable the user to specify different time ranges for when the data was (or is) relevant in a search (Priority = 3)		
F-21	**Provide user accounts, with storage and retrieval of data and workflows all within a user's account**	2	F-21a - Provide user accounts (Priority = 2)	F-21b - storage and retrieval of data within a user's account (Priority = 2)	F-21c - Record workflows within a user's account (Priority = 2)
F-22	**Locate information related to the topic in question, but restricted to specified countries**	2	F-22a - restrict searches to specified countries (Priority = 2)		
F-23	**Provide a facility to assess the impact of a policy decision**	2	F-23a - Provide a facility to assess the impact of a policy decision (Priority = 2)		

Original Requirement			Atomic Decomposition		
F-ID	**Description**	**Pri-ority**	**A**	**B**	**C**
F-24	**Ability to use data from both the European Commission and the European Parliament**	2	F-24a - Ability to use data from the European Commission (Priority = 2)	F-24b - Ability to use data from the European Parliament (Priority = 2)	
F-25	**Augment search terms the user has specified and perform additional and related searches to the ones the user is running**	1	F-25a - (the system should) Augment search terms the user has specified (Priority = 1)	F-25b - (the system should) perform automatic additional and related searches to the ones the user is running (Priority = 1)	
F-26	**Support multiple languages, in the user interface labelling, searching and analyses**	1	F-26a - Support multiple languages in searches (Priority = 1)	F-26b - Support multiple languages in user interface labelling (Priority = 1)	F-26c - Support multiple languages in analyses (Priority = 1)
F-27	**Provide summaries or commentaries of raw data sets**	3	F-27a - Provide summaries of raw data sets (Priority = 3)	F-27b - Provide commentaries of raw data sets (Priority = 3)	
F-28	**Visualisation of data sets**	3	F-28a - Visualisation of data sets (Priority = 3)		
F-29	**Highlight any data from social media**	3	F-29a - Highlight any data from social media (Priority = 3)		
F-30	**The data origin should be a filtration criterion**	3	F-30a - Enable filtration of data (Priority = 3)	F-30b - Filter by data origin (Priority = 3)	
F-31	**Determination of whether and how multiple data sets can be compared**	3	F-31a - Determine if two data sets are comparable (Priority = 3)	F-31b - Determination of whether and how multiple data sets can be compared (Priority = 3)	
F-32	**Find documents and data sets related to the policy element in question from countries outside the native country of the user**	2	F-32a - Enable searches for data from countries outside the user's own (Priority = 2)		
F-33	**Present a summary of a search result to the user**	2	F-33a - Present a summary of a search result to the user (Priority = 2)		
F-34	**Present available metadata about a search result to the user**	2	F-34a - Present available metadata about a search result to the user (Priority = 2)		

Original Requirement			Atomic Decomposition		
F-ID	**Description**	**Pri-ority**	**A**	**B**	**C**
F-35	**Enable the user to select which metadata they frequently want to see in their search result summaries**	2	F-35a - Enable the user to select which metadata they frequently want to see in their search result summaries (Priority = 2)		
F-36	**Locate discussions on social media that are related to the policy issue in question**	2	F-36a - Locate discussions on social media that are related to the policy issue in question (Priority = 2)		
F-37	**Enable the user to specify a date that determines the earliest creation point of data in search results**	1	F-37a - Enable the user to specify a date that determines the earliest creation point of data in search results (Priority = 1)		
F-38	**In the event of multiple versions of a data set being available, highlight the most recent**	1	F-38a - In the event of multiple versions of a data set being available, highlight the most recent (Priority = 1)		
F-39	**Match an impacted factor with a government department**	1	F-39a - Match an impacted factor with a government department (Priority = 1)		
F-40	**Ability for the policy maker user to customise the tool to their preferences**	1	F-40a - Ability for the policy maker user to customise the tool to their preferences (Priority = 1)		
F-41	**The tool's UI and the analyses in the research partners' components support German as a language**	3	F-41a - The tool's UI should support German as a language (Priority = 3)	F-41b - The analysis components should support German as a language (Priority = 3)	
F-42	**Construction of searches based on the policy themes in question**	3	F-42a - Construction of searches based on the policy themes in question (Priority = 3)		
F-43	**Ranking of search results using different criteria**	3	F-43a - Ranking of search results using different criteria (Priority = 3)		

Original Requirement			Atomic Decomposition		
F-ID	**Description**	**Pri-ority**	**A**	**B**	**C**
F-44	**Monitor blogs and comments on them**	2	F-44a - Monitor blogs (Priority = 2)	F-44b - Monitor comments on blogs (Priority = 2)	

Table 2: Atomic Functional Requirements

The reasons for the requirements highlighted in red that will not be addressed are discussed next. For F-4, the reason for this is that we cannot collect data from traditional online media because the content itself is subject to copyright. In general, the comments on the traditional media websites are subject to "database rights", which means we cannot access the comments freely. In addition, we must abide by the terms and conditions of the site. Access to sites such as this is also problematic. In general, social media sites have official APIs that provide access to the postings, but most traditional online media does not. Given that we will not screen scrape (this is usually a violation of the terms and conditions of accessing the sites), we cannot collect data from any site that does not have an official API. The same argument applies to F-44, blogs. Most blog hosting platforms are subject to database right, and do not have official APIs. The remaining requirement that will not be addressed is F-39 "match a factor with a government department". This is simply not necessary – we assert that anyone working for government should have a very good idea of which department a factor should be applicable to. The remaining requirement, F-27b "Provide commentaries of raw data sets" will not be addressed because commentaries are usually subjective opinions, and while subjective opinions from social media are useful and it is clearly understood that they may have some kind of bias, to have commentaries on data sets, where any bias is much less clear, is not useful.

4.2.2 Organise Functional Requirements

Once we had a definitive set of atomic requirements (Table 2), it became apparent that many were related. The major relationship we discovered was that some requirements were more specific versions of others. This enabled us to organise the requirements into a hierarchy of generality. We were also able to identify broad functional groups, corresponding to the overall key concepts of the project. This was seen as a good result, since the requirements from end users were in broad agreement with the overall project concepts. The result of this analysis is a diagram of the functional requirements where collections of specific requirements are gathered under the general heading of a more general requirement. The diagram is broken up into a series of smaller diagrams (Figure 9 to Figure 13 inclusive), corresponding to the broad functional groups, and each is discussed separately in the following pages.

Each diagram concentrates on a specific major topic and is generally organised with the specific requirements below the more general requirements that group the specific requirements together. The arrows in the diagram connect specific requirements to their

more general relations, and the direction of the arrow points to the more general requirement.

The diagrams also have colour coding, and the meaning of this is as follows.

- A green node means that there is no more specific requirement of the same type. These have been so-called “leaf nodes” in our analysis, as they represent the most specific requirements of a type.
- A blue node denotes a requirement where there exist more specific requirements of the same type. Blue nodes represent broad functional themes - they typically collect together the set of more specific requirements below.
- A grey requirement means that is made redundant by another requirement, and the two requirements are connected together to illustrate that they are similar and one is made redundant.
- A red requirement is not addressed (the discussion following Table 2 above gives reasons).

From the hierarchical structure in the diagrams illustrating the general / more specific requirements of the same type, we have determined a principle that determines which requirements need addressing:

The green (leaf node) requirements are the ones that need to be addressed. This is because they are the most specific of their class. The blue requirements above them in the hierarchy represent broad functional themes. Within these functional themes, the green nodes determine the specific requirements.

As a result of this analysis, we have a definitive set of specific requirements that need to be addressed by the architecture (those that are green in the figures).

4.2.2.1 Search Requirements

Figure 9 shows the requirements, grouped for search-related functionality. It can be seen that there are some major groups in the figure, and these correspond broadly to some of the DOW functions, which provides confidence in the approach of the project. These broad groups are as follows.

- F-1b - Search multiple data sources
- F-42a - Construction of searches based on the policy themes in question
- F-7a - Enable finding information for subjects related to those in the policy
- F-15b - Enable the collection of opinions about the policy subjects

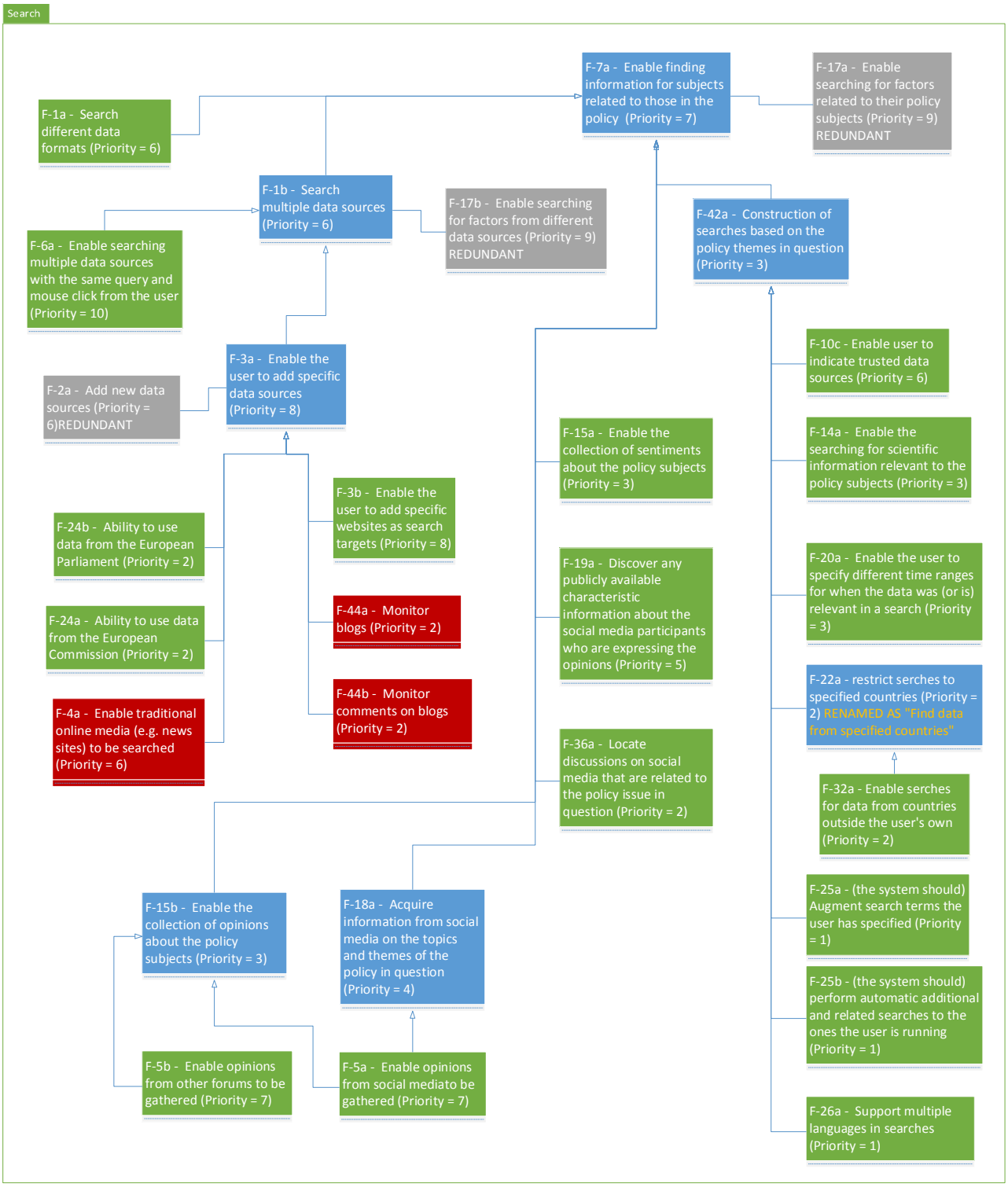


Figure 9: Search Requirements Hierarchy

4.2.2.2 Result Presentation Requirements

Figure 10 shows the requirements concerned with the presentation of results. The main themes of these requirements are:

- F-9a - Show all available provenance information for a piece of data
- F-8a - Enable the user to filter search results (include or exclude results according to a wide range of criteria)
- F-11a - Enable the user to sort data by different criteria
- F-33a - Present a summary of a search result to the user

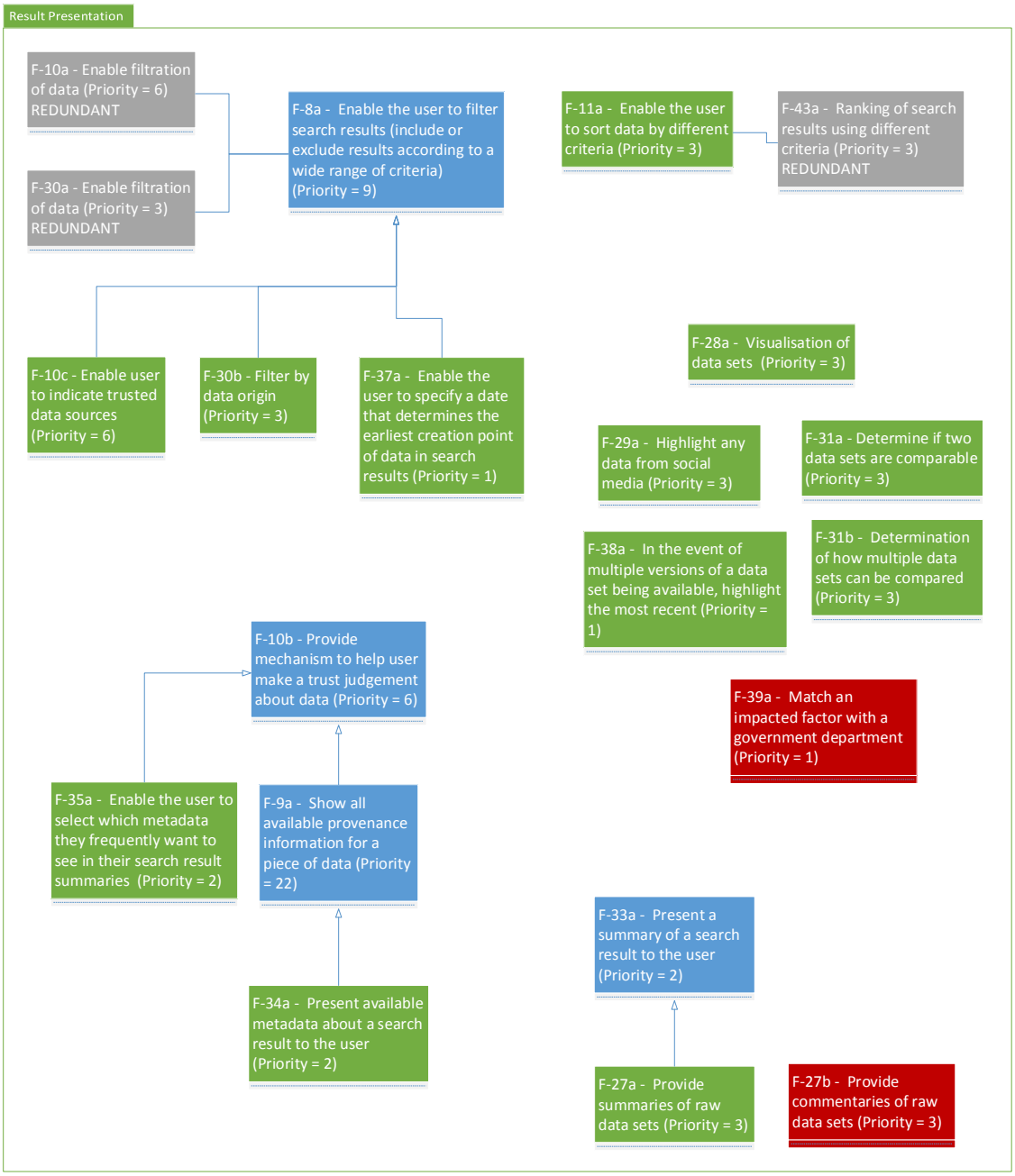


Figure 10: Result Presentation Requirement Hierarchy

4.2.2.3 Policy Impact Assessment Requirements

Figure 11 shows the requirements concerned with policy impact assessment. In this section, the requirements are sparser than the previously-discussed requirement groups, and this is attributable to the level of understanding the policy maker end users have of this subject – it is highly likely to be new to them, so it is unlikely that they have strong opinions or requirements for it. There is only one group in this section, but the main themes within this group are as follows.

- F13-a and F-13b – Transparency in both creation and operation of policy models
- F-16b - The policy maker should be able to add new factors to policy model

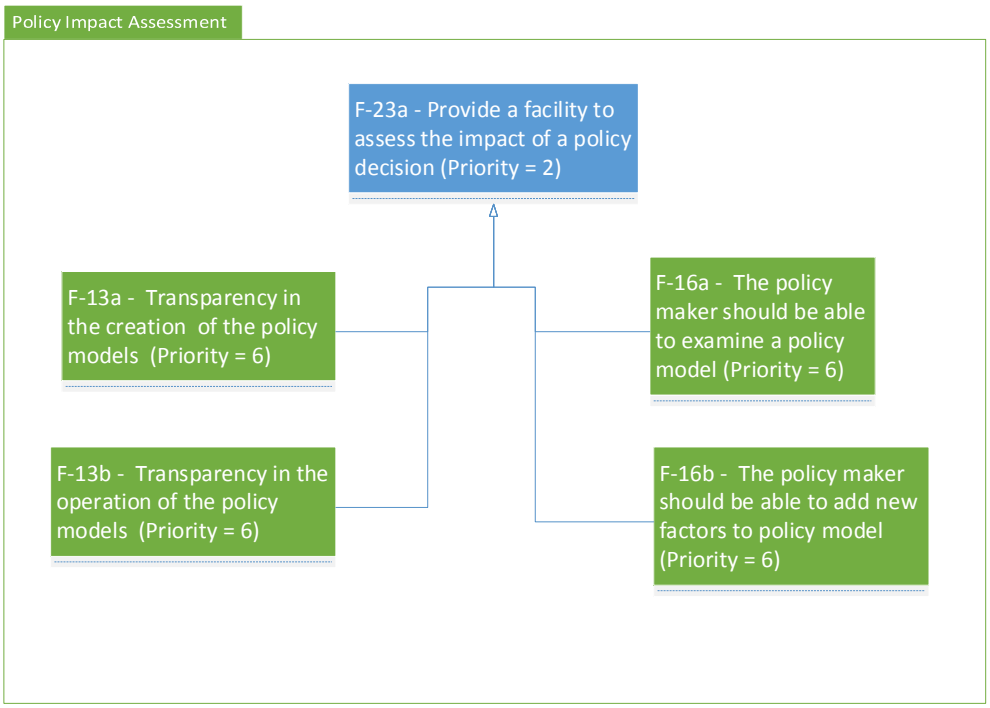


Figure 11: Policy Impact Assessment Functional Requirements Hierarchy

4.2.2.4 User Accounts & History Requirements

Figure 12 shows the requirements concerned with user accounts & recording of history. There is only one group here, but within this (apart from the usability aspect of user customisation of the toolkit), there are two major requirement groups, as follows.

- Provenance represented by recording of audit trails: F-21c - Record workflows within a user's account; and F-9b - (Show history of what user has done with data inside the toolkit).
- F-21b - storage and retrieval of data within a user's account.

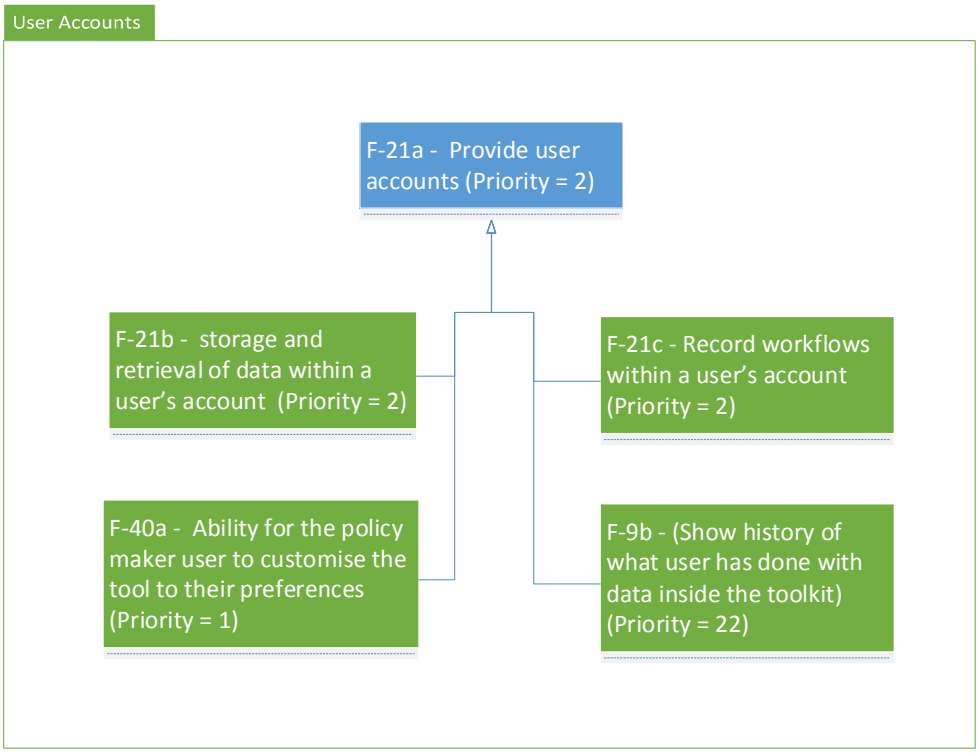


Figure 12: User Accounts Functional Requirements Hierarchy

4.2.2.5 Language Support Requirements

Figure 13 shows the requirements connected to multiple language support. The two groups are clear:

- F-26c - Support multiple languages in analyses
- F-26b - Support multiple languages in user interface labelling

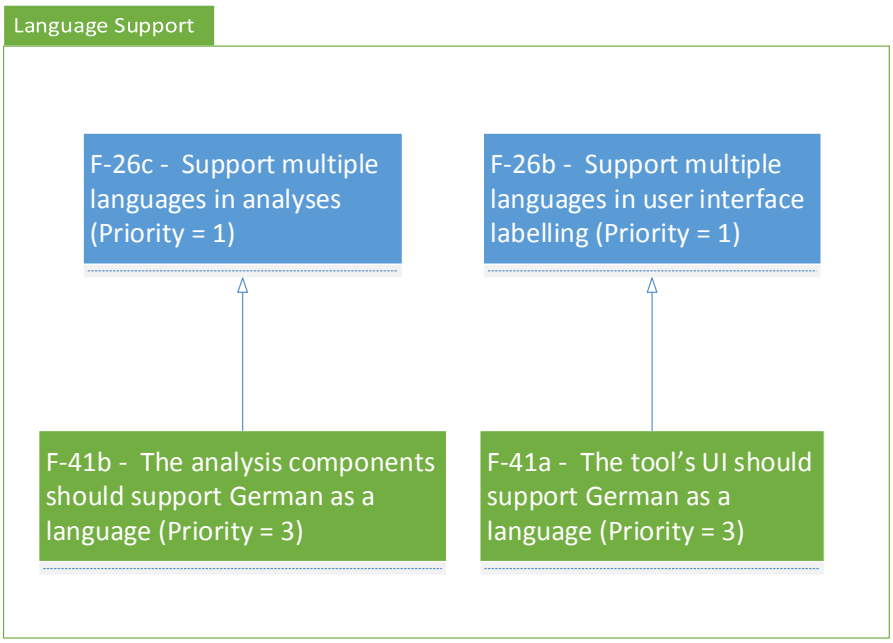


Figure 13: Language Support Functional Requirements Hierarchy

4.2.3 Determine How to Address Requirements

The analysis above determined a master set of base, "leaf node", requirements (the green ones in the diagrams), which are the ones that should be addressed. These will be addressed, to a greater or lesser extent. Comments on how each requirement is addressed are in the following table These comments are derived from the comments about how to address requirements in D3.1, but also updated based on progress and greater understanding since the time D3.1 was written. For each requirement, a discussion is provided on how the project will address it, and to what extent.

F-ID	Description	Pri-ority	How to Address
F-1a	Search different data formats	6	Currently LOD and Social media searches are supported via the LOD search and social media search tools. These are integrated into the system via wrappers and controllers, and all data found is put into the database. Additional formats will be investigated and added as necessary.

F-ID	Description	Pri-ority	How to Address
F-3b	Enable the user to add specific websites as search targets	8	A box will be provided to allow URLs of data sources and websites to be added manually by the user, to be included in a group of searches (see response to F-6). Data sources can take many forms, from a URL of a linked open data site, to the twitter account of a social media user. Ideally it should be possible to specify these different types of data source, but the project will need to investigate which types are most effective to support. It is unlikely every type of data source will be supported. There are likely to be classes of data source (e.g. straightforward web page, linked open data, social media), and some of these can be preset into lists from which the user can select. Users may have favourite data source starting points, so it should also be possible to store these in a user's profile.
F-5a	Enable opinions from social media to be gathered	7	The social media analysis tools developed in Sense4us will compute sentiment and opinions from user generated content and provide these insights to the policy makers.
F-5b	Enable opinions from other forums to be gathered	7	This may be supported, but is highly dependent on the forums selected. Only forums with data access APIs may be supported, and many forums do not support API access.
F-6a	Enable searching multiple data sources with the same query and mouse click from the user	10	Users will be able to specify search terms, or use the output of another Sense4us tool as search terms, and then be able to specify the sources of data to search (e.g. social media, open data etc). All searches conducted as a result of this group of searches will have their results stored separately in the Sense4us database under the user's profile, but they will be also accessible as a group corresponding to the group of searches.
F-9b	Show history of what user has done with data inside the toolkit	22	This will be supported by the UI and the infrastructure by showing a graph of data the user has collected or the outputs of analysis, along with the history path of how that data was generated or collected, showing e.g. which analysis tool was used, what its settings were, which data source was used, what the search terms were, etc.
F-10c	Enable user to indicate trusted data sources	6	The user will be presented with whatever metadata is available about data sources. From this they can make a judgement about which they trust, and can then mark those data source as trusted.

F-ID	Description	Pri-ority	How to Address
F-11a	Enable the user to sort data by different criteria	3	Results should be able to be sorted, and the criteria is likely to be many and varied. Results will be sort-able where possible, and this depends on the criteria. It is easy to sort numerically, alphabetically or chronologically, but other types of criteria will be examined on a case by case basis to determine whether it is possible. Where it is not possible to sort, the UI will indicate this by greying out any "sort" buttons or tools.
F-13a	Transparency in the creation of the policy models	6	This will be achieved in a number of ways: by providing tools to help a user build a model for the policy problem, by structuring the problem in terms of the key variables and parameters identified by the user; allowing the user to define and modify links, policy objectives in terms of targeted changes in outcome indicators; and providing documentation of the methodology used for the simulation tool.
F-13b	Transparency in the operation of the policy models	6	The simulation can be re-run a number of times, to enable the user to test different scenarios, and by adjustment of one parameter at a time, will enable them to determine the effect of that parameter.
F-14a	Enable the searching for scientific information relevant to the policy subjects	3	This will be addressed as far as possible, and will depend on the data sources available. "Scientific information" can take a number of forms, e.g. peer-refereed publications, data sets from experiments, open data from public studies etc. Where possible, these will be made available, but will depend on the availability and compatibility of the data sources.
F-15a	Enable the collection of sentiments about the policy subjects	3	It is most likely that the sentiments about the policy subjects will be derived from social media, as these are places where people make such comments. It is also possible to collect data from many social media forums, as they have official APIs, so it is possible to collect the opinions whilst still remaining compliant with their terms and conditions and privacy regulations.
F-16a	The policy maker should be able to examine a policy model	6	The policy model is intended to be represented as an open graph of nodes and links, so the policy maker can examine its interior.

F-ID	Description	Pri-ority	How to Address
F-16b	The policy maker should be able to add new factors to policy model	6	The policy model is intended to be represented as an open graph of nodes and links, so the policy maker can examine its interior and identify any gaps. The other tools (e.g. the searches and analyses) are intended to help the policy maker find more related information, so these tools can be used to add nodes and links to the policy model graph.
F-19a	Discover any publicly available characteristic information about the social media participants who are expressing the opinions	5	To address this requirement the tool can provide all user data publicly available in the social media platforms (e.g., name of the user, location, description, time in the platform, etc.) as well as additional insights of those users based on their activities in the platform (post rate, size of their social network, etc.). Even though this data is public, it still can identify individual people, so is regarded as personal data, and therefore data protection applies. Compliance with data protection entails significant work, so the project needs to decide whether this requirement is significant enough to warrant the effort. The project also needs to be careful here to avoid any perception of surveillance of citizens.
F-20a	Enable the user to specify different time ranges for when the data was (or is) relevant in a search	3	This can be addressed in two ways: restricting the search time range before the search is executed (i.e. as parameters to the search request), or filtration of search results. In either case, restricting search results by date is only possible when items inside data sources have date information, so therefore some searches may not be possible to date-restrict.
F-21b	storage and retrieval of data within a user's account	2	All results from searches and analyses are stored in a back-end database, and are stored connected with the user account that executed the search or analysis.
F-21c	Record workflows within a user's account	2	The processing a user does will be recorded, so that they have access to a history of the data and its transformations. This is useful for provenance - showing the user how they got to a particular result for example. It is also important to enable the user access to intermediate data within their history of processing, and this is supported by recording all data items the user has collected or generated though analysis.
F-24a	Ability to use data from the European	2	This can be supported as long as data is available and with supported APIs to access it.

F-ID	Description	Pri-ority	How to Address
	Commission		
F-24b	Ability to use data from the European Parliament	2	This can be supported as long as data is available and with supported APIs to access it.
F-25a	(The system should) Augment search terms the user has specified	1	Synonyms for the user's search terms can be found for search terms using lexicon lookups such as finding owl:sameAs links discovered from some registry (e.g., sameas.org). Usually when accessing data about an entity owl:sameAs toward other data sources are provided. For instance, the entity dbpedia:Germany (in the DBpedia sata source) has an owl:sameAs with fb:Germany (in the Freebase data source).
F-25b	(The system should) perform automatic additional and related searches to the ones the user is running	1	Once synonyms are found from a user's search terms (F-25a), additional searches can be automatically set running and the results can be collected with the results from the user's original search.
F-26a	Support multiple languages in searches	1	This will be supported if the search APIs of the data sources support it. We will not support all languages within the project, but consider the ones needed directly for the end user partners.
F-27a	Provide summaries of raw data sets	3	Where possible, the tool will display summaries of data sets, but it is not possible to guarantee that all data sets will be able to be summarised. Some data sets will come with summaries, and these can be shown to the user (if they are correctly identified). In other cases, Sense4us may be able to provide automatically generated summaries of data (e.g. from documents). In general, each data type and source will need to be examined on a case by case basis to determine whether a summary is present and how it can be identified, or if a summary can be generated.
F-28a	Visualisation of data sets	3	This requirement is very general, as each data set is different and will have different visualisation criteria. In addition, there are many different visualisation types, and some are more suited to some data sets than others. Also, different users prefer to see things in different ways. Each data type encountered will be examined on a case by case basis to determine if it can be visualised, and how it can be visualised. This is potentially a vast task and needs to be made tractable, so it would be a good idea to evolve a standard set of

F-ID	Description	Pri-ority	How to Address
			visualisations, based on what is popular with end users, what is possible given the tools available, and what is compatible with the data.
F-29a	Highlight any data from social media	3	This is simply a case of showing the data source to the user and pointing out that it is social media.
F-30b	Filter by data origin	3	The user will be shown the data sources from which data came, and will be able to select those that should be shown in results.
F-31a	Determine if two data sets are comparable	3	It is expected that if this requirement is addressed at all, it will be addressed minimally and under tightly constrained conditions. Whether it is possible to address this requirement depends strongly on the data being compared and its compatibility. It is unlikely that automatic detection of compatible data sets will be possible, due to the multitudes of different data formats and semantics that may be encountered. Having said this, some fields may be possible to compare ("date" for example).
F-31b	Determination of how multiple data sets can be compared	3	It is expected that if this requirement is addressed at all, it will be addressed minimally and under tightly constrained conditions. Whether it is possible to address this requirement depends strongly on the data being compared and its compatibility. It is unlikely that automatic detection of compatible data sets will be possible, due to the multitudes of different data formats and semantics that may be encountered. Having said this, some fields may be possible to compare ("date" for example).
F-32a	Enable searches for data from countries outside the user's own	2	This will be supported where the APIs of the data sources support it. For example, some social media search tools can specify countries and locations to restrict a search to.
F-34a	Present available metadata about a search result to the user	2	All available metadata about a search result will be recorded in the database alongside the search results themselves, and so the user can see any of this metadata.
F-35a	Enable the user to select which metadata they frequently want to see in their search result summaries	2	The user profile will contain a section for favourite metadata, and this allows the user to select metadata that will be presented in search results.

F-ID	Description	Pri-ority	How to Address
F-36a	Locate discussions on social media that are related to the policy issue in question	2	The location of information will be based on the selection of keywords representing the policy. These keywords can be voluntarily provided by the users or automatically extracted from a policy document.
F-37a	Enable the user to specify a date that determines the earliest creation point of data in search results	1	If time of publishing of the dataset is available, the user will have the option to set the earliest date as a filtration criteria.
F-38a	In the event of multiple versions of a data set being available, highlight the most recent	1	This can be addressed by ranking the same dataset by creation date, so that the most recent is at the top.
F-40a	Ability for the policy maker user to customise the tool to their preferences	1	User profiles with customisation functions will be provided. There are many other functional requirement responses that contain customisation and user profile aspects (e.g. favourite filters and trusted source lists), and these will be gathered together to provide the specification for the user profile section of the Sense4us toolkit.
F-41a	The tool's UI should support German as a language	3	The UI will be developed so that it may be skinned with different languages and a translation file supplied for German.
F-41b	The analysis components should support German as a language	3	Searches can be conducted using German data sources and with German language keywords. For analysis tools, English Natural Language Processing (NLP) libraries will be replaced by German NLP libraries to enable analysing data in both languages.

Table 3: How to Address Functional Requirements

4.3 DOW Functional Sub-Architectures

This section contains a scenario and associated sub-architecture for each of the relevant DOW functions. Each scenario description includes:

- the steps the user is likely to go through,
- the input and output,
- components involved, and
- relevant functional requirements.

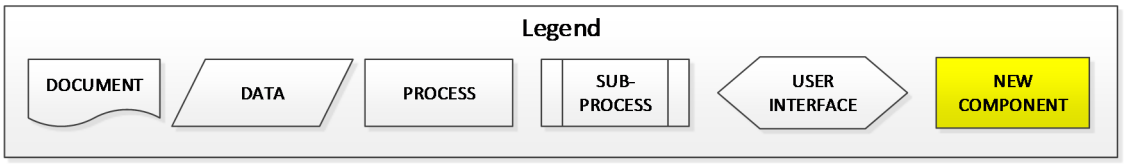
Each scenario also includes a sub-architecture picture with a brief discussion that shows the components required and how they interact in order to address the functional scenario. Once all the sub-architectures are determined, they are combined into the overall architecture for Sense4us.

Because the relationships between the DOW functions described in Figure 7 where one DOW function is useful for another, the DOW functional scenarios will be described in the order of dependency with independent ones first. This is so that each one progressively builds on ones already described.

Some notes on the DOW function scenarios follow:

- The diagrams describing the sub-architectures may contain duplicated components, for example there may be two databases or two UIs in a diagram. This is simply for clarity in the diagrams – there is only one database and UI in the architecture.
- New components that have not been previously identified (i.e. they were not imagined in the Initial Proposition) are highlighted in yellow in the diagrams.
- The relevant requirements are included for each scenario. Each table is a cut-down version of Table 3 (how to address functional requirements) containing the requirements that are relevant to the DOW function scenario. The whole row for each relevant requirement is repeated from Table 3, for the purposes of completeness so that each scenario is self-standing.

For reference, the diagrams use the following legend:



4.3.1 DOW-FN-16 Theme Analysis

This scenario covers extraction of keywords from source texts (documents or tweets are currently supported, but other types may be supported in the future).

DOW-Function	Description
DOW Function ID	DOW-FN-16
Name	Extract Themes from Documents
Purpose	Documents or other data is summarised with keywords describing it.
Input	Documents OR Tweets (info from social media) (Other data may be supported in future)
Input DOW-Function(s)	-

Output	Keywords summarising input
Consumer DOW-Function(s)	Useful for DOW-FN-15 Generation of Search Terms
Operation Sequence	1. User uploads documents or selects from previously collected tweets 2. User runs analysis: a. Analysis produces summary keywords b. Keywords are saved in DB c. Provenance record is produced and saved in DB 3. User can view results and provenance information stored in DB

Table 4: Theme Analysis Description

F-ID	Description	Pri-ority	How to Address
F-9b	Show history of what user has done with data inside the toolkit	22	This will be supported by the UI and the infrastructure by showing a graph of data the user has collected or the outputs of analysis, along with the history path of how that data was generated or collected, showing e.g. which analysis tool was used, what its settings were, which data source was used, what the search terms were, etc.
F-27a	Provide summaries of raw data sets	3	Where possible, the tool will display summaries of data sets, but it is not possible to guarantee that all data sets will be able to be summarised. Some data sets will come with summaries, and these can be shown to the user (if they are correctly identified). In other cases, Sense4us may be able to provide automatically generated summaries of data (e.g. from documents). In general, each data type and source will need to be examined on a case by case basis to determine whether a summary is present and how it can be identified, or if a summary can be generated.
F-41b	The analysis components should support German as a language	3	Searches can be conducted using German data sources and with German language keywords. For analysis tools, English Natural Language Processing (NLP) libraries will be replaced by German NLP libraries to enable analysing data in both languages.

Table 5: Relevant Functional Requirements for Theme Analysis

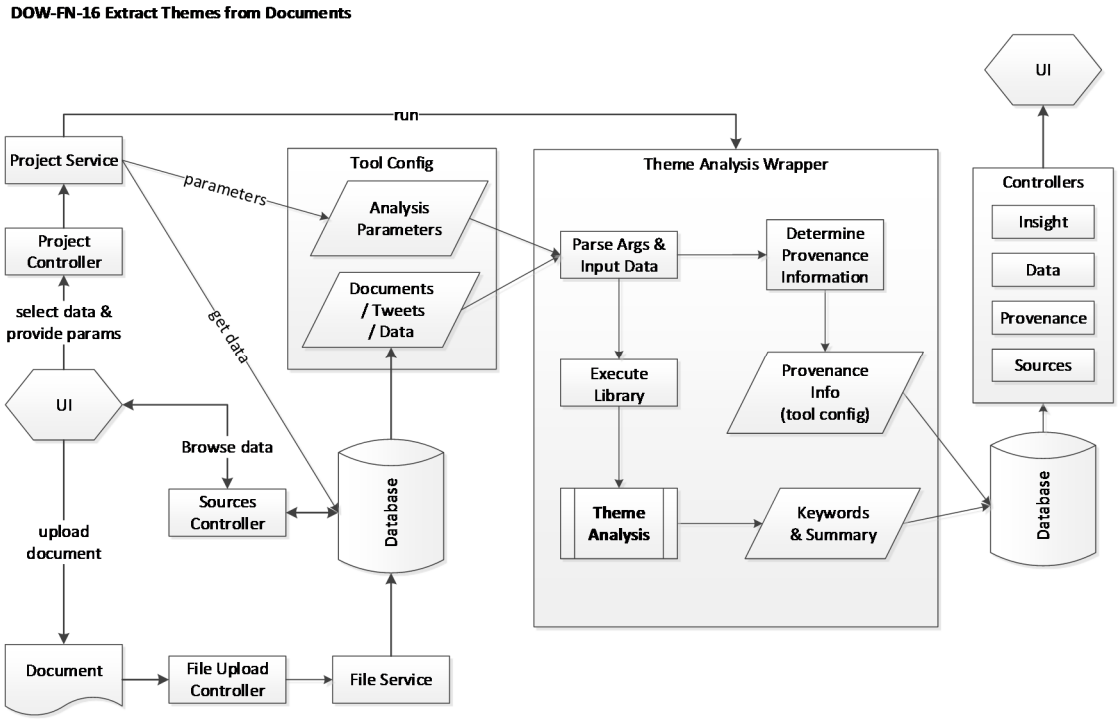


Figure 14: Theme Analysis Sub-Architecture

Figure 14 shows the components and sub-architecture for theme analysis. The user interacts with the UI. The user can optionally upload a document using the File Upload Controller, which invokes the File Service to store the file in the DB. The user can then select a document they have uploaded or other textual data that is stored in the DB. The user can do this by browsing the data using the Sources Controller, and selecting data items they want analysed. The selected data items are passed via the UI and Project Controller to the Project Service, which prepares the input parameters, gets the relevant data from the database and passes it to the Theme Analysis Wrapper. The Project Service then launches the Theme Analysis Wrapper, which parses the arguments and executes the Theme Analysis library. The Theme Analysis library itself is a piece of code supplied by U Koblenz and generates the keywords summarising the input data. The wrapper generates provenance information consisting of the complete set of input parameters, a reference to the input data, analysis type and version, who ran the analysis, when they ran it etc, and this is stored in the database along with the output of the analysis. The user can then view a snippet of the output using the Source Controller, or the full output with the Insight Controller.

4.3.2 DOW-FN-15 Generation of Search Terms

This scenario covers the generation of search terms, which are used in subsequent functions (such as social media search). This scenario looks for similar search terms to those already supplied, so as to augment them, to give the user further suggestions.

DOW-Function	Description
DOW Function ID	DOW-FN-15
Name	Generation of Search Terms

Purpose	Finds additional search terms similar or related to the ones supplied
Input	Keywords describing search terms
Input DOW-Function(s)	(optionally) DOW-FN-2 Find Related Information
Output	Keywords or URIs describing related or similar search terms
Consumer DOW-Function(s)	Useful for DOW-FN-1 Intelligent Search of Heterogeneous Sources and DOW-FN-8 Social Media Search
Operation Sequence	1. User specifies keywords or selects keywords from previous analyses 2. User runs analysis: a. Analysis produces related keywords b. Keywords are saved in DB c. Provenance record is produced and saved in DB 3. User can view results and provenance information stored in DB

Table 6: Generation of Search Terms Description

F-ID	Description	Pri-ority	How to Address
F-9b	Show history of what user has done with data inside the toolkit	22	This will be supported by the UI and the infrastructure by showing a graph of data the user has collected or the outputs of analysis, along with the history path of how that data was generated or collected, showing e.g. which analysis tool was used, what its settings were, which data source was used, what the search terms were, etc.
F-25a	(The system should) Augment search terms the user has specified	1	Synonyms for the user's search terms can be found for search terms using lexicon lookups such as finding owl:sameAs links discovered from some registry (e.g., sameas.org). Usually when accessing data about an entity owl:sameAs toward other data sources are provided. For instance, the entity dbpedia:Germany (in the DBpedia sata source) has an owl:sameAs with fb:Germany (in the Freebase data source).
F-41b	The analysis components should support German as a language	3	Searches can be conducted using German data sources and with German language keywords. For analysis tools, English Natural Language Processing (NLP) libraries will be replaced by German NLP libraries to enable analysing data in both languages.

Table 7: Functional Requirements Relevant for Generation of Search Terms

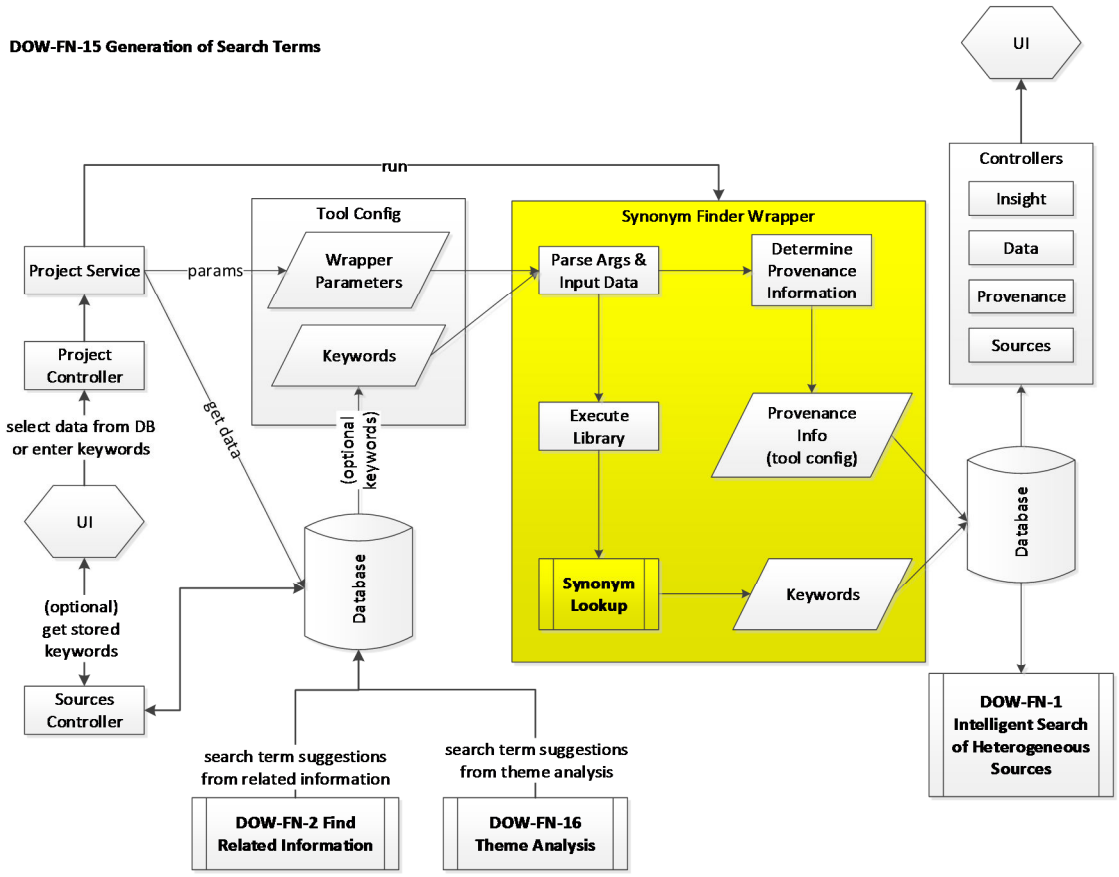


Figure 15: Generation of Search Terms Sub-Architecture

Figure 15 shows the components and sub-architecture for generation of search terms. The user interacts with the UI, and they can select keywords from previous analyses (stored in the DB) or type some keywords in. The user can select previous analysis results by browsing the data using the Sources Controller, and selecting data items they want analysed. The selected data items are passed via the UI and Project Controller to the Project Service, which prepares the input parameters, gets the relevant data from the database and passes it to the Synonym Finder Wrapper (a new component, highlighted in yellow). The Project Service then launches the Synonym Finder Wrapper, which parses the arguments and executes the Synonym Lookup library. The Synonym Lookup library itself is a piece of code supplied by UKoblenz that looks up synonyms and other items from external registries such as http://sameas.org/ and returns synonyms or related search terms compared to the input data. The wrapper generates provenance information consisting of the complete set of input parameters, a reference to the input data, analysis type and version, who ran the analysis, when they ran it etc, and this is stored in the database along with the output of the analysis. The user can then view a snippet of the output using the Source Controller, or the full output with the Insight Controller.

4.3.3 DOW-FN-8 Social Media Search

This scenario covers searching of social media sites for postings related to the supplied search terms. Currently Twitter is supported and other sites may be supported later if there is a requirement for this.

DOW-Function	Description
DOW Function ID	DOW-FN-8
Name	Social Media Search
Purpose	Search, retrieve and store social media postings relevant to the search terms supplied.
Input	Search terms in the form of keywords
Input DOW-Function(s)	DOW-FN-15 Generation of Search Terms
Output	Social media postings
Consumer DOW-Function(s)	Useful for DOW-FN-1 Intelligent Search of Heterogeneous Sources Needed for DOW-FN-9 Social Media Discussion Analysis
Operation Sequence	1. User selects keywords from previous analyses or types keywords in 2. User specifies data sources (currently fixed to Twitter) and search parameters, .e.g. how far back in time to search. 3. User runs search: a. Tweets returned from Twitter API b. Tweets are saved in DB c. Provenance record is produced and saved in DB 4. User can view results and provenance information stored in DB

Table 8: Social Media Search Description

F-ID	Description	Pri-ority	How to Address
F-1a	Search different data formats	6	Currently LOD and Social media searches are supported via the LOD search and social media search tools. These are integrated into the system via wrappers and controllers, and all data found is put into the database. Additional formats will be investigated and added as necessary.
F-3b	Enable the user to add specific websites as search targets	8	A box will be provided to allow URLs of data sources and websites to be added manually by the user, to be included in a group of searches (see response to F-6). Data sources can take many forms, from a URL

			of a linked open data site, to the twitter account of a social media user. Ideally it should be possible to specify these different types of data source, but the project will need to investigate which types are most effective to support. It is unlikely every type of data source will be supported. There are likely to be classes of data source (e.g. straightforward web page, linked open data, social media), and some of these can be present into lists from which the user can select. Users may have favourite data source starting points, so it should also be possible to store these in a user's profile.
F-5a	Enable opinions from social media to be gathered	7	The social media analysis tools developed in Sense4us will compute sentiment and opinions from user generated content and provide these insights to the policy makers.
F-9b	Show history of what user has done with data inside the toolkit	22	This will be supported by the UI and the infrastructure by showing a graph of data the user has collected or the outputs of analysis, along with the history path of how that data was generated or collected, showing e.g. which analysis tool was used, what its settings were, which data source was used, what the search terms were, etc.
F-15a	Enable the collection of sentiments about the policy subjects	3	It is most likely that the sentiments about the policy subjects will be derived from social media, as these are places where people make such comments. It is also possible to collect data from many social media forums, as they have official APIs, so it is possible to collect the opinions whilst still remaining compliant with their terms and conditions and privacy regulations.
F-19a	Discover any publicly available characteristic information about the social media participants who are expressing the opinions	5	To address this requirement the tool can provide all user data publicly available in the social media platforms (e.g., name of the user, location, description, time in the platform, etc.) as well as additional insights of those users based on their activities in the platform (post rate, size of their social network, etc.). Even though this data is public, it still can identify individual people, so is regarded as personal data, and therefore data protection applies. Compliance with data protection entails significant work, so the project needs to decide whether this requirement is significant enough to warrant the effort. The project also needs to be careful here to avoid any perception of surveillance of citizens.
F-	Enable the user to specify different time	3	This can be addressed in two ways: restricting the search time range before the search is executed (i.e.

20a	ranges for when the data was (or is) relevant in a search		as parameters to the search request), or filtration of search results. In either case, restricting search results by date is only possible when items inside data sources have date information, so therefore some searches may not be possible to date-restrict.
F-21b	storage and retrieval of data within a user's account	2	All results from searches and analyses are stored in a back-end database, and are stored connected with the user account that executed the search or analysis.
F-25b	(The system should) perform automatic additional and related searches to the ones the user is running	1	Once synonyms are found from a user's search terms (F-25a), additional searches can be automatically set running and the results can be collected with the results from the user's original search.
F-26a	Support multiple languages in searches	1	This will be supported if the search APIs of the data sources support it. We will not support all languages within the project, but consider the ones needed directly for the end user partners.
F-32a	Enable searches for data from countries outside the user's own	2	This will be supported where the APIs of the data sources support it. For example, some social media search tools can specify countries and locations to restrict a search to.
F-34a	Present available metadata about a search result to the user	2	All available metadata about a search result will be recorded in the database alongside the search results themselves, and so the user can see any of this metadata.
F-36a	Locate discussions on social media that are related to the policy issue in question	2	The location of information will be based on the selection of keywords representing the policy. These keywords can be voluntarily provided by the users or automatically extracted from a policy document.
F-41b	The analysis components should support German as a language	3	Searches can be conducted using German data sources and with German language keywords. For analysis tools, English Natural Language Processing (NLP) libraries will be replaced by German NLP libraries to enable analysing data in both languages.

Table 9: Relevant Functional Requirements for Social Media Search

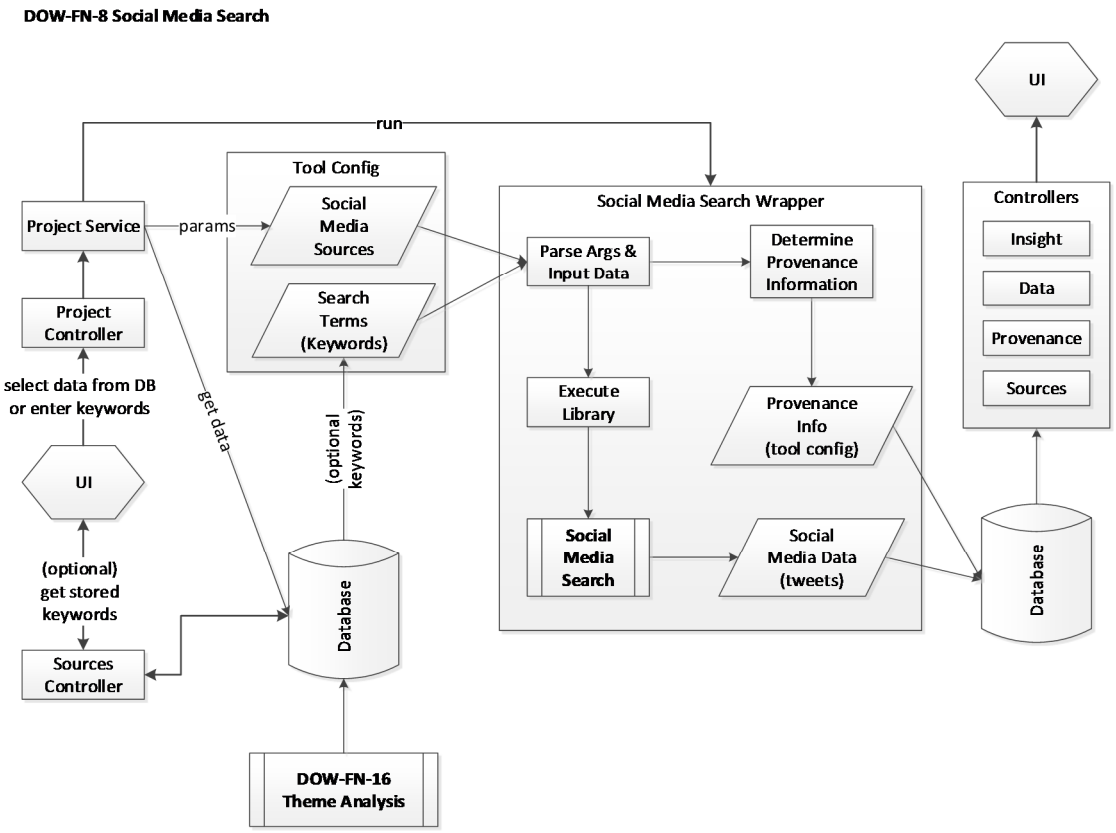


Figure 16: Social Media Search Sub-Architecture

Figure 16 shows the components and sub-architecture for social media search. The user interacts with the UI, and they can select keywords from previous analyses (stored in the DB) or type some keywords in to use as search terms. The selected data items are passed via the UI and Project Controller to the Project Service, which prepares the input parameters, gets any relevant keywords from the database and passes it to the Social Media Search Wrapper. The Project Service then launches the Social Media Search Wrapper, which parses the arguments and executes the Social Media Search library. The Social Media Search library itself is client to the official Twitter API that takes parameters as a query and returns tweets in JSON format. The wrapper generates provenance information consisting of the complete set of input parameters, a reference to the input data, analysis type and version, who ran the analysis, when they ran it etc, and this is stored in the database along with the search results. The user can then view a snippet of the output using the Source Controller, or the full output with the Data Controller.

4.3.4 DOW-FN-1 Intelligent Search of Heterogeneous Sources

This scenario has the purpose of orchestrating together searches from multiple data sources.

DOW-Function	Description
DOW Function ID	DOW-FN-1
Name	Intelligent Search of Heterogeneous Sources
Purpose	Multiple different data sources are searched in a single

	search.
Input	Search Terms
Input DOW-Function(s)	DOW-FN-15 Generation of Search Terms DOW-FN-8 Social Media Search
Output	Results of different data types
Consumer DOW-Function(s)	Useful for DOW-FN-2 Find Related Information
Operation Sequence	1. User selects keywords from previous analyses or types keywords in 2. User specifies data sources and search parameters. 3. User runs search: a. Tweets returned from Twitter API b. Tweets are saved in DB c. LOD search is ran and LOD graph generated. d. LOD graph saved in DB e. Provenance record is produced and saved in DB 4. User can view results and provenance information stored in DB

Table 10: Search of Heterogeneous Sources Description

F-ID	Description	Pri-ority	How to Address
F-1a	Search different data formats	6	Currently LOD and Social media searches are supported via the LOD search and social media search tools. These are integrated into the system via wrappers and controllers, and all data found is put into the database. Additional formats will be investigated and added as necessary.
F-3b	Enable the user to add specific websites as search targets	8	A box will be provided to allow URLs of data sources and websites to be added manually by the user, to be included in a group of searches (see response to F-6). Data sources can take many forms, from a URL of a linked open data site, to the twitter account of a social media user. Ideally it should be possible to specify these different types of data source, but the project will need to investigate which types are most effective to support. It is unlikely every type of data source will be supported. There are likely to be classes of data source (e.g. straightforward web page, linked open data, social media), and some of these can be preset into lists from which the user can select. Users may have favourite data source starting points, so it should also be possible to store these in a user's profile.

F-6a	Enable searching multiple data sources with the same query and mouse click from the user	10	Users will be able to specify search terms, or use the output of another Sense4us tool as search terms, and then be able to specify the sources of data to search (e.g social media, open data etc). All searches conducted as a result of this group of searches will have their results stored separately in the Sense4us database under the user's profile, but they will be also accessible as a group corresponding to the group of searches.
F-15a	Enable the collection of sentiments about the policy subjects	3	It is most likely that the sentiments about the policy subjects will be derived from social media, as these are places where people make such comments. It is also possible to collect data from many social media forums, as they have official APIs, so it is possible to collect the opinions whilst still remaining compliant with their terms and conditions and privacy regulations.
F-20a	Enable the user to specify different time ranges for when the data was (or is) relevant in a search	3	This can be addressed in two ways: restricting the search time range before the search is executed (i.e. as parameters to the search request), or filtration of search results. In either case, restricting search results by date is only possible when items inside data sources have date information, so therefore some searches may not be possible to date-restrict.
F-24a	Ability to use data from the European Commission	2	This can be supported as long as data is available and with supported APIs to access it.
F-24b	Ability to use data from the European Parliament	2	This can be supported as long as data is available and with supported APIs to access it.
F-25b	(The system should) perform automatic additional and related searches to the ones the user is running	1	Once synonyms are found from a user's search terms (F-25a), additional searches can be automatically set running and the results can be collected with the results from the user's original search.
F-26a	Support multiple languages in searches	1	This will be supported if the search APIs of the data sources support it. We will not support all languages within the project, but consider the ones needed directly for the end user partners.
F-36a	Locate discussions on social media that are related to the policy issue in question	2	The location of information will be based on the selection of keywords representing the policy. These keywords can be voluntarily provided by the users or automatically extracted from a policy document.
F-37a	Enable the user to specify a date that determines the earliest creation point of data in	1	If time of publishing of the dataset is available, the user will have the option to set the earliest date as a filtration criteria.

	search results		

Table 11: Relevant Functional Requirements for Search of Heterogeneous Sources

The sub-architecture for searching heterogeneous data sources is a combination of two searches, currently: DOW-FN-8, Social Media Search, and a new search that needs to be defined: DOW-FN-1a, Linked Open Data Search. Before discussing the detail of the search of heterogeneous source, the LOD Search sub-case will be described. This is shown in Figure 17.

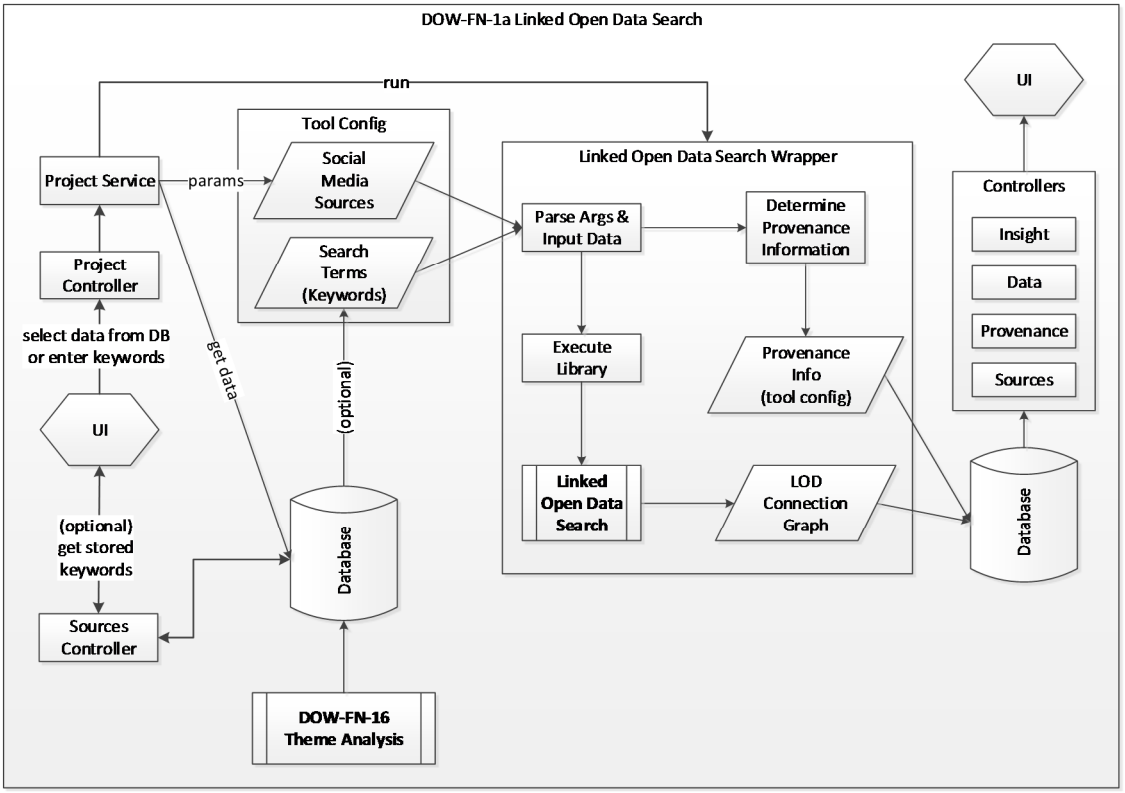


Figure 17: LOD Search Sub-Architecture

The LOD search is very similar to the Social Media search, in that it uses the same actions outside the LOD Search Wrapper to set up the search – the user can choose to use original keywords or ones stored in the DB from previous analyses. The Project Service runs the LOD Search Wrapper, which runs the LOD search library. This is supplied by UKoblenz and discovers resources in LOD and how they are linked together. The result is a graph of LOD resources and connections, which is stored in the database along with the standard provenance information.

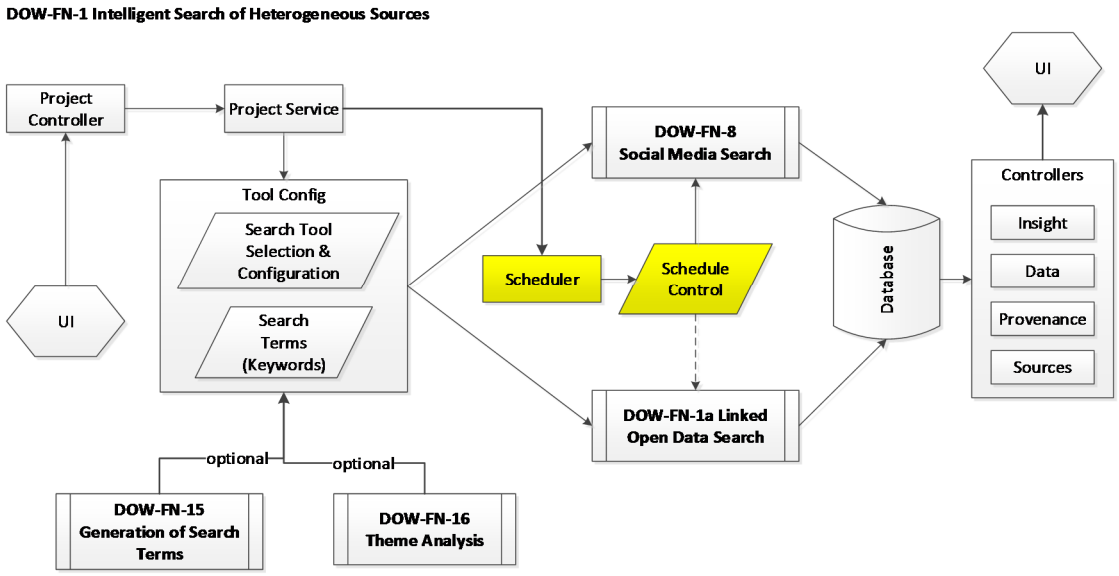


Figure 18: Search of Heterogeneous Sources Sub-Architecture

The complete sub architecture for the search of heterogeneous source is shown in Figure 18. This combines the social media search from DOW-FN-8 and the LOD search from DOW-FN-1a. The usage pattern is the same as for individual searches, and this sub-architecture adds the orchestration of multiple searches so the user can search these two different data sources with one overall search.

The scheduler and its output (the schedule control) are new components, and their purpose is to execute searches on a schedule. This may be a repeating search or a long-term search for example. The scheduler is most useful when it controls the social media search, executing the same search repeated at regular intervals for example. It may be the case that the scheduler executes repeated Linked Open Data searches, but this is less likely than with social media because social media changes much more quickly than the data in Linked Open Data repositories and repeated searches are therefore more useful with social media sources to catch the latest comments. This is why there is a solid connection from the schedule control to the social media search and a dashed connection to Linked Open Data search.

The pattern is extensible to other searches if there is a requirement to add searching other data types. For each new search needed, a new wrapper will need to be written (using the existing wrappers as templates and guides) to support all the specific aspects of the new search and translate them to standard APIs supported by components such as the DB and the Project Service.

4.3.5 DOW-FN-13 Provenance Information Extraction and Display

This scenario covers extraction of provenance information from all other scenarios. In effect, it is a summary of the work done in each other scenario to collect and support provenance information. Some of the source scenarios have not yet been introduced, but the pattern for extracting provenance information is the same for all.

The analysis of creating the system architecture with the associated analysis of functional requirements has shown that in general, there are three major types of provenance relevant to Sense4us:

- Tool run configuration – this is information about one run of a Sense4us tool. It can record information such as the type of tool (e.g. the name and version of the library executed by the wrapper), input parameters to the tool, input data to the tool (this can be an identifier to other data stored within the Sense4us database, who ran the tool, the date and time the tool was ran, etc.
- Metadata about the data from external sources – when data is collected from an external source, any available metadata is recorded about the name of the data source, the author of the data set, the date the data set was created, etc.
- Records of the user's decisions, actions and selections – when the user makes decisions (for example selecting some subset of data to be input to another tool), this needs to be recorded so that there is a history of what the user did, what data they used, and how they used it. This is related to the tool run configuration – if a user took some subset of data from the output of one tool, added some information of their own, and put the result into another tool for analysis, this needs to be recorded.

DOW-Function	Description
DOW Function ID	DOW-FN-13
Name	Provenance Information Extraction and Display
Purpose	Record provenance information from user actions and metadata, and display it
Input	-
Input DOW-Function(s)	Any of: • DOW-FN-1a Linked Open Data Search • DOW-FN-2 Find Related Information • DOW-FN-5 Policy Model Construction • DOW-FN-6 Simulation of Policy Option Impact • DOW-FN-7 Policy Validation in Context of Previous Similar Policy • DOW-FN-8 Social Media Search • DOW-FN-9 Social Media Discussion Analysis • DOW-FN-14 Evaluation of Multiple Policy Options • DOW-FN-15 Generation of Search Terms • DOW-FN-16 Theme Analysis • DOW-FN-17 Evidence Extraction
Output	Provenance information stored in database
Consumer DOW-Function(s)	Needed for DOW-FN-17 Evidence Extraction
Operation Sequence	1. User runs one of the input scenarios 2. Inside the application wrapper, available information

	is gathered concerning: a. Application configuration (e.g. arguments, parameters and metadata) b. User actions and selections c. Any applicable metadata from the data source d. Any applicable metadata from the results of the input scenario 3. Provenance information is stored in the DB along with the run that generated it. 4. User can view results and provenance information stored in DB

Table 12: Provenance Extraction Description

F-ID	Description	Pri-ority	How to Address
F-9b	Show history of what user has done with data inside the toolkit	22	This will be supported by the UI and the infrastructure by showing a graph of data the user has collected or the outputs of analysis, along with the history path of how that data was generated or collected, showing e.g. which analysis tool was used, what its settings were, which data source was used, what the search terms were, etc.
F-10c	Enable user to indicate trusted data sources	6	The user will be presented with whatever metadata is available about data sources. From this they can make a judgement about which they trust, and can then mark those data source as trusted.
F-21c	Record workflows within a user's account	2	The processing a user does will be recorded, so that they have access to a history of the data and its transformations. This is useful for provenance - showing the user how they got to a particular result for example. It is also important to enable the user access to intermediate data within their history of processing, and this is supported by recording all data items the user has collected or generated though analysis.
F-34a	Present available metadata about a search result to the user	2	All available metadata about a search result will be recorded in the database alongside the search results themselves, and so the user can see any of this metadata.

Table 13: Relevant Functional Requirements for Provenance Extraction

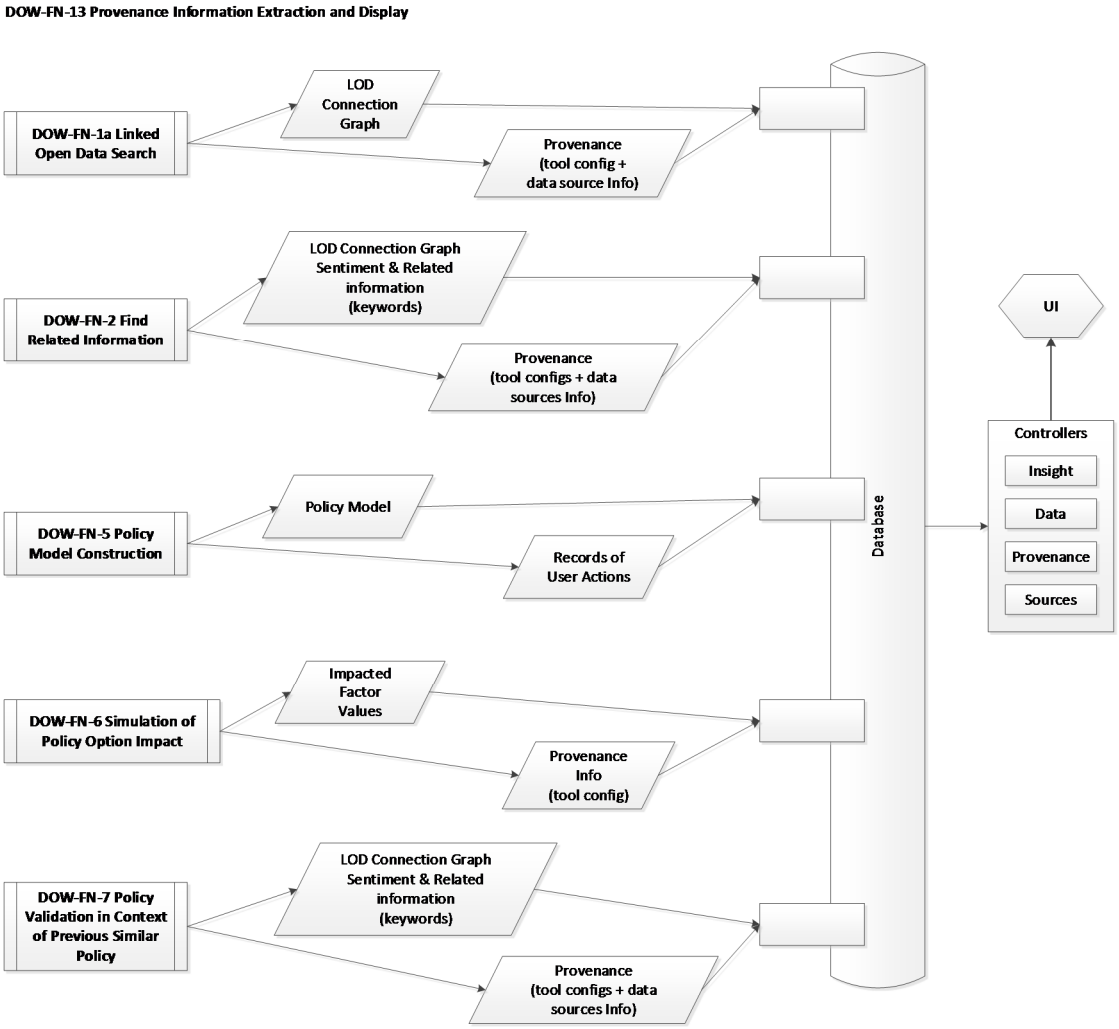


Figure 19: Provenance Extraction Sub-Architecture (part 1)

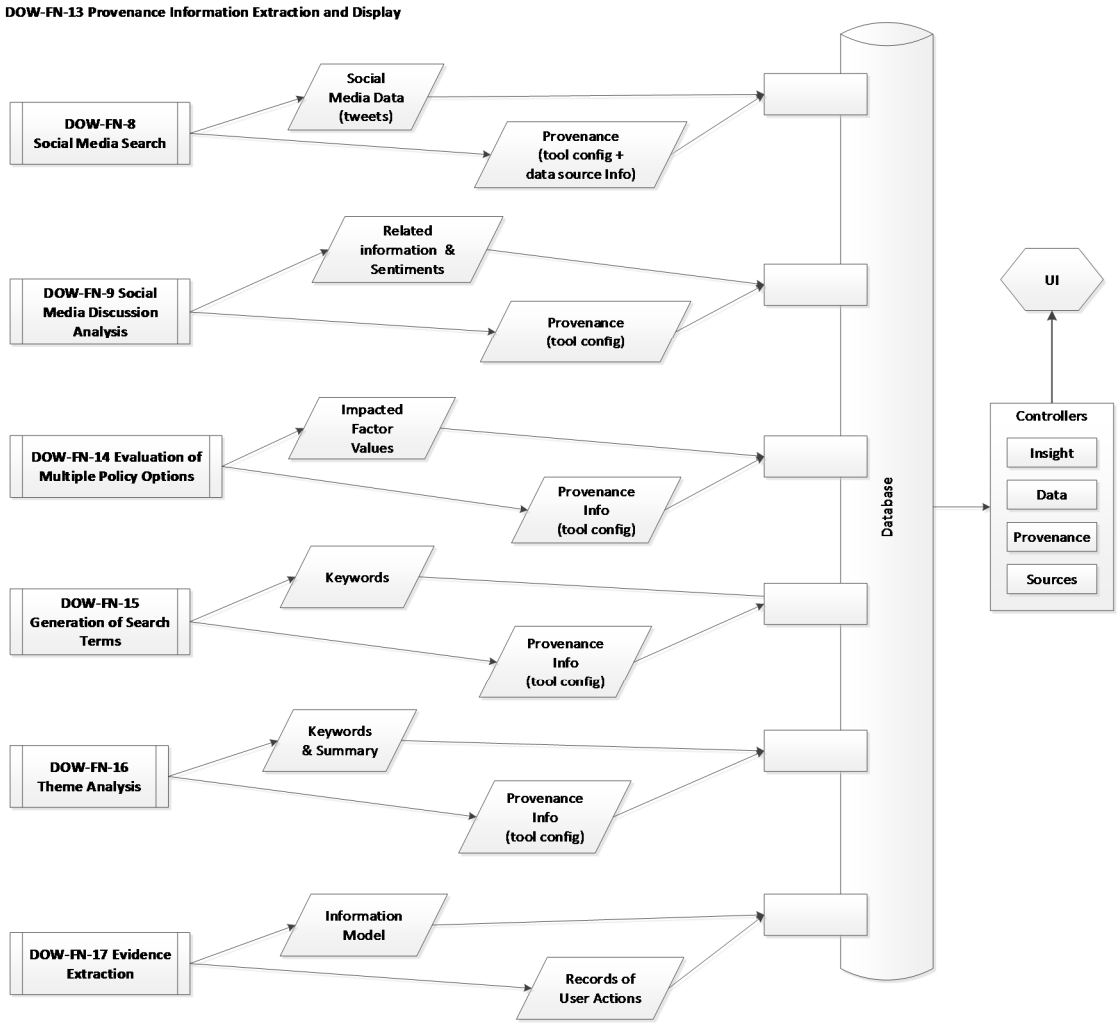


Figure 20: Provenance Extraction Sub-Architecture (part 2)

The sub-architecture for provenance extraction is shown split into Figure 19 and Figure 20 for reasons of space. Each DOW functional scenario that generates provenance information provides slightly different information, but there are many common factors across all the different information.

4.3.6 DOW-FN-9 Social Media Discussion Analysis

This scenario covers the analysis of social media discussions using the tools supplied by KMi.

DOW-Function	Description
DOW Function ID	DOW-FN-9
Name	Social Media Discussion Analysis
Purpose	Discover sentiment and related information to the policy subject areas from social media postings.
Input	Tweets (info from social media) (Other social media data may be supported in future)

Input DOW-Function(s)	DOW-FN-8 Social Media Search
Output	For each word in the data set, a set of related keyword and distances is created, together with a measure of the sentiment of that word.
Consumer DOW-Function(s)	Needed for DOW-FN-2 Find Related Information
Operation Sequence	1. User selects from previously collected tweets 2. User runs analysis: a. Analysis produces sets of words and related words and sentiments b. Results are saved in DB c. Provenance record is produced and saved in DB 3. User can view results and provenance information stored in DB

Table 14: Social Media Discussion Analysis Description

F-ID	Description	Pri-ority	How to Address
F-5a	Enable opinions from social media to be gathered	7	The social media analysis tools developed in Sense4us will compute sentiment and opinions from user generated content and provide these insights to the policy makers.
F-5b	Enable opinions from other forums to be gathered	7	This may be supported, but is highly dependent on the forums selected. Only forums with data access APIs may be supported, and many forums do not support API access.
F-15a	Enable the collection of sentiments about the policy subjects	3	It is most likely that the sentiments about the policy subjects will be derived from social media, as these are places where people make such comments. It is also possible to collect data from many social media forums, as they have official APIs, so it is possible to collect the opinions whilst still remaining compliant with their terms and conditions and privacy regulations.
F-19a	Discover any publicly available characteristic information about the social media participants who are expressing the opinions	5	To address this requirement the tool can provide all user data publicly available in the social media platforms (e.g., name of the user, location, description, time in the platform, etc.) as well as additional insights of those users based on their activities in the platform (post rate, size of their social network, etc.). Even though this data is public, it still can identify individual people, so is regarded as personal data, and therefore data protection applies. Compliance with data protection entails significant work, so the project needs to decide

			whether this requirement is significant enough to warrant the effort. The project also needs to be careful here to avoid any perception of surveillance of citizens.
F-36a	Locate discussions on social media that are related to the policy issue in question	2	The location of information will be based on the selection of keywords representing the policy. These keywords can be voluntarily provided by the users or automatically extracted from a policy document.
F-41b	The analysis components should support German as a language	3	Searches can be conducted using German data sources and with German language keywords. For analysis tools, English Natural Language Processing (NLP) libraries will be replaced by German NLP libraries to enable analysing data in both languages.

Table 15: Relevant Functional Requirements for Social Media Discussion Analysis

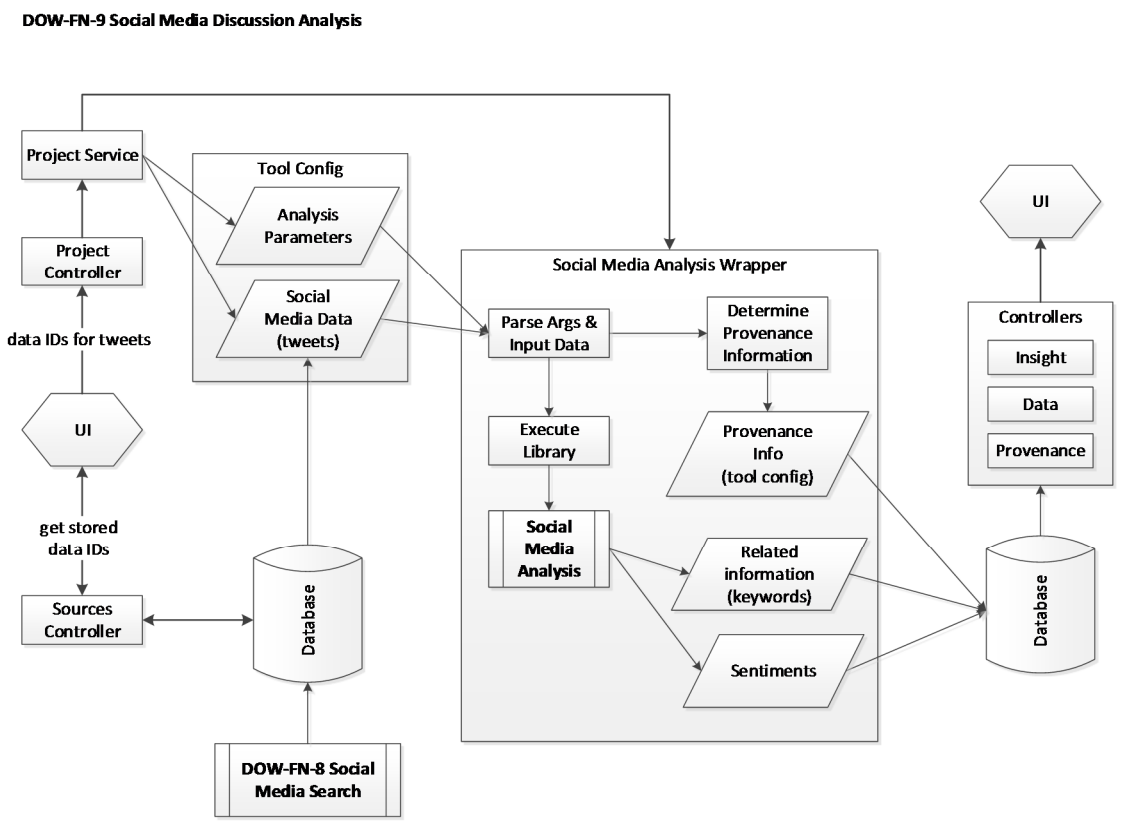


Figure 21: Social Media Discussion Analysis Sub-Architecture

The social media analysis is integrated into the Sense4us infrastructure in the same way that other tools are – via a wrapper. The social media takes tweets as input and this is selected by the user from previously-run social media searches. Once the analysis has run, the output and provenance information are stored in the database by the wrapper.

4.3.7 DOW-FN-2 Find Related Information

This scenario covers an important area for the Sense4us toolkit – helping the user find information in and around their policy area, by suggesting concepts, entities, stakeholders and organisations that they may not be aware of.

DOW-Function	Description
DOW Function ID	DOW-FN-2
Name	Find Related Information
Purpose	Give the user a set of information related to a core input term, and show how it is related.
Input	Search terms in the form of keywords
Input DOW-Function(s)	DOW-FN-9 Social Media Discussion Analysis DOW-FN-1 Intelligent Search of Heterogeneous Sources
Output	Related information in the form of related terms from social media and a connected graph of concepts from LOD.
Consumer DOW-Function(s)	Needed for DOW-FN-17 Evidence Extraction Useful for DOW-FN-15 Generation of Search Terms Useful for DOW-FN-5 Policy Model Construction Useful for DOW-FN-7 Policy Validation in Context of Previous Similar Policy
Operation Sequence	1. User inputs or selects keywords from previous analyses 2. User runs analysis: a. Analysis produces related information b. Results are saved in DB c. Provenance record is produced and saved in DB 3. User can view results and provenance information stored in DB

Table 16: Find Related Information

F-ID	Description	Pri-ority	How to Address
F-5a	Enable opinions from social media to be gathered	7	The social media analysis tools developed in Sense4us will compute sentiment and opinions from user generated content and provide these insights to the policy makers.
F-6a	Enable searching multiple data sources with the same query and mouse click from the	10	Users will be able to specify search terms, or use the output of another Sense4us tool as search terms, and then be able to specify the sources of data to search (e.g social media, open data etc). All searches

	user		conducted as a result of this group of searches will have their results stored separately in the Sense4us database under the user's profile, but they will be also accessible as a group corresponding to the group of searches.
F-14a	Enable the searching for scientific information relevant to the policy subjects	3	This will be addressed as far as possible, and will depend on the data sources available. "Scientific information" can take a number of forms, e.g. peer-refereed publications, data sets from experiments, open data from public studies etc. Where possible, these will be made available, but will depend on the availability and compatibility of the data sources.
F-15a	Enable the collection of sentiments about the policy subjects	3	It is most likely that the sentiments about the policy subjects will be derived from social media, as these are places where people make such comments. It is also possible to collect data from many social media forums, as they have official APIs, so it is possible to collect the opinions whilst still remaining compliant with their terms and conditions and privacy regulations.
F-25a	(The system should) Augment search terms the user has specified	1	Synonyms for the user's search terms can be found for search terms using lexicon lookups such as finding owl:sameAs links discovered from some registry (e.g., sameas.org). Usually when accessing data about an entity owl:sameAs toward other data sources are provided. For instance, the entity dbpedia:Germany (in the DBpedia sata source) has an owl:sameAs with fb:Germany (in the Freebase data source).
F-25b	(The system should) perform automatic additional and related searches to the ones the user is running	1	Once synonyms are found from a user's search terms (F-25a), additional searches can be automatically set running and the results can be collected with the results from the user's original search.
F-36a	Locate discussions on social media that are related to the policy issue in question	2	The location of information will be based on the selection of keywords representing the policy. These keywords can be voluntarily provided by the users or automatically extracted from a policy document.
F-41b	The analysis components should support German as a language	3	Searches can be conducted using German data sources and with German language keywords. For analysis tools, English Natural Language Processing (NLP) libraries will be replaced by German NLP libraries to enable analysing data in both languages.

Table 17: Relevant Functional Requirements for Find Related Information

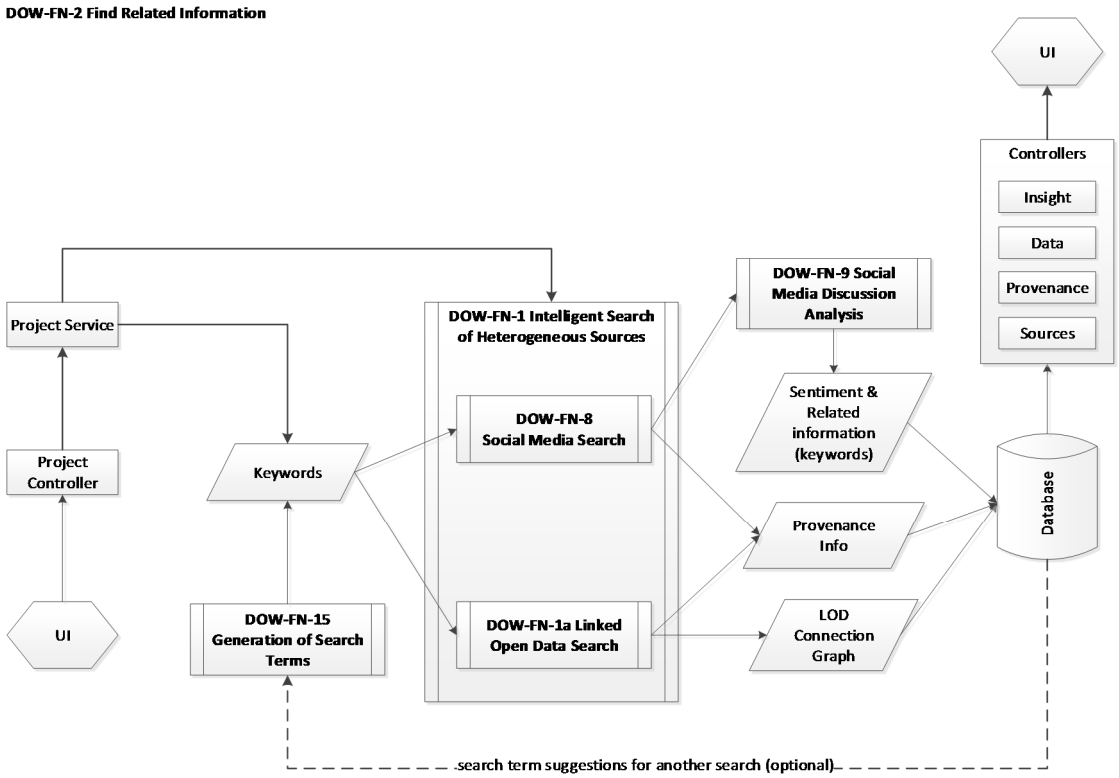


Figure 22: Find Related Information Sub-Architecture

Figure 22 shows the sub-architecture for the find related information case. This is essentially the heterogeneous search case (DOW-FN-1) followed by social media discussion analysis (DOW-FN-9). This is because the output of these two DOW functions provides information that is related to input keywords.

There is a "feedback loop" in this case where the related information output can also be used as inputs for search term generation (DOW-FN-15). This illustrates that the tools can be used in a number of different orders, and that the user is in ultimate control of the use the tools are put to. Here, the example is that the user could discover some interesting related information, and they want to dig further into this, so they use this to ask the tool to generate some new related search terms and begin the cycle again.

4.3.8 DOW-FN-17 Evidence Extraction

This scenario aims to help the user build an information model. The intention of the information model is to provide a place where the user can record pieces of information they deem important, and how these pieces of information are related (if they are related). The information model is a precursor to the policy model constructed in DOW-FN-5 Policy Model Construction, and the main difference between the two is that the information model contains all kinds of relationships, whereas the policy model is more concerned with relationships that represent cause and effect. Hence the policy model can be regarded as a subset of the information model. The information model is useful in its own right, as it helps the user organise the information they have gathered through the tool.

There is no set workflow in the building of an information model – the user can go to other tools and find more information, and incorporate this into the information model at any time. As such, the information model acts as a kind of scratch pad where the user can record the important information they have discovered.

DOW-Function	Description
DOW Function ID	DOW-FN-17
Name	Evidence Extraction
Purpose	Find related information along with provenance information, and organise that information so its relationships and connections can be seen.
Input	Related information keywords, keyword distances, URIs, LOD connectivity graph
Input DOW-Function(s)	DOW-FN-2 Find Related Information
Output	Information model built by the user showing concepts and relationships
Consumer DOW-Function(s)	Useful for DOW-FN-5 Policy Model Construction Useful for DOW-FN-7 Policy Validation in Context of Previous Similar Policy
Operation Sequence	1. User has previously gathered related information from DOW-FN-2. Provenance records for the related information serves as part of the evidence. 2. User can select from this information and paste the selection into an "information model" - a place to record concepts and relationships. 3. User can return to other analyses and add information to their information model at any time.

Table 18: Evidence Extraction Description

F-ID	Description	Pri-ority	How to Address
F-9b	Show history of what user has done with data inside the toolkit	22	This will be supported by the UI and the infrastructure by showing a graph of data the user has collected or the outputs of analysis, along with the history path of how that data was generated or collected, showing e.g. which analysis tool was used, what its settings were, which data source was used, what the search terms were, etc.
F-13a	Transparency in the creation of the policy models	6	This will be achieved in a number of ways: by providing tools to help a user build a model for the policy problem, by structuring the problem in terms of the key variables and parameters identified by the user; allowing the user to define and modify

			links, policy objectives in terms of targeted changes in outcome indicators; and providing documentation of the methodology used for the simulation tool.
F-16a	The policy maker should be able to examine a policy model	6	The policy model is intended to be represented as an open graph of nodes and links, so the policy maker can examine its interior.
F-16b	The policy maker should be able to add new factors to policy model	6	The policy model is intended to be represented as an open graph of nodes and links, so the policy maker can examine its interior and identify any gaps. The other tools (e.g. the searches and analyses) are intended to help the policy maker find more related information, so these tools can be used to add nodes and links to the policy model graph.

Table 19: Relevant Functional Requirements for Evidence Extraction

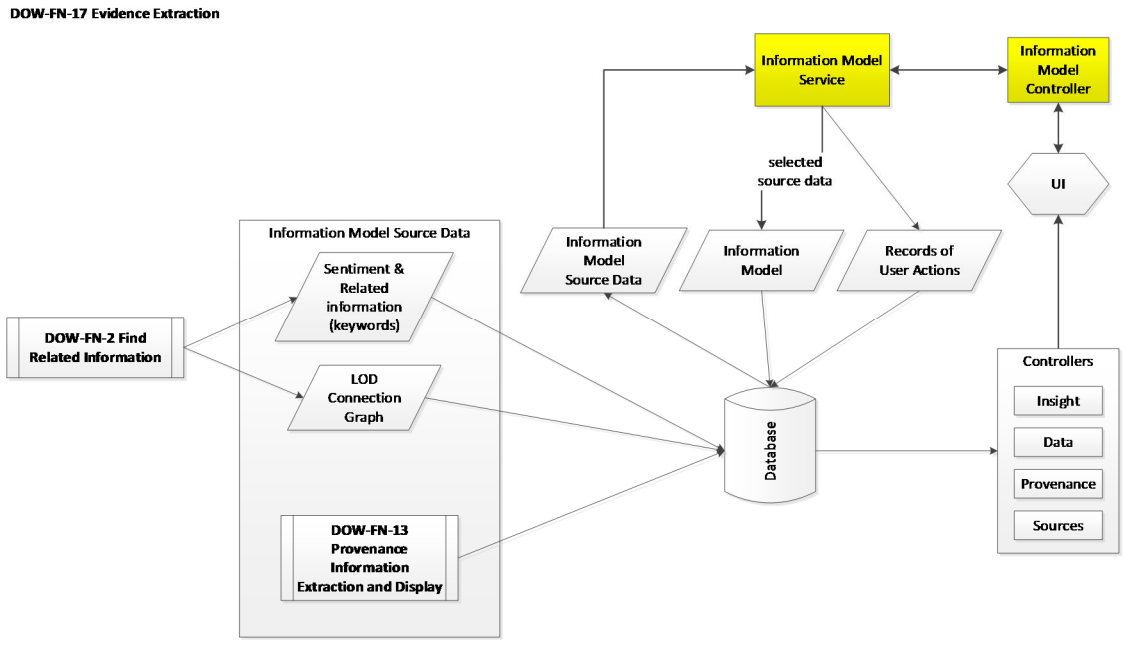


Figure 23: Evidence Extraction Sub-Architecture

Figure 23 shows the sub-architecture for evidence extraction. The user interacts with the UI and the Information Model Controller and the Information Model Service does the main work. These are both new components planned after the D7.1 demonstrator. The user can select elements from DOW-FN-2, find related information, as elements in the information model. The elements selected are entirely up to the user – DOW-FN-2 simply provides source material in the form of related information keywords, URIs and LOD connection graphs, along with the associated provenance information. All this comes from the Sense4us database, where it was stored upon previous collection or analysis. The user can add to or delete from

the information model at any time, and can use any analysis output they want as source data for the information model. A history of the actions the user has done on their information model is recorded, so that the user can go back to a previous version if they so wish. This is especially useful if the user has deleted something from their model by mistake or they change their mind.

4.3.9 DOW-FN-7 Policy Validation in Context of Previous Similar Policy

This scenario covers validation of policy aspects by comparison with similar previous policies. This is done by running similar analyses (represented by DOW-FN-2) on aspects from previous policies, e.g. reaction to its aspects on social media, and comparing them to the current policy being evaluated.

DOW-Function	Description
DOW Function ID	DOW-FN-7
Name	Policy Validation in Context of Previous Similar Policy
Purpose	Compare related information derived from previous policies to the related information derived from a current policy.
Input	Document(s) describing previous policies Document(s) describing current policy
Input DOW-Function(s)	DOW-FN-2 Find Related Information is useful
Output	Sets of related information from each policy (previous and current)
Consumer DOW-Function(s)	-
Operation Sequence	1. User uploads documents or selects from previously collected tweets 2. User runs analysis: a. Analysis produces summary keywords b. Keywords are saved in DB c. Provenance record is produced and saved in DB 3. User can view results and provenance information stored in DB

Table 20: Policy Validation in Context of Previous Similar Policy

This scenario covers no functional requirements in addition to those already covered by DOW-FN-2 (find related information). This is mainly because this scenario is really a way of using DOW-FN-2 to compare the related information derived from different policy documents (i.e. previous ones and the current one under investigation).

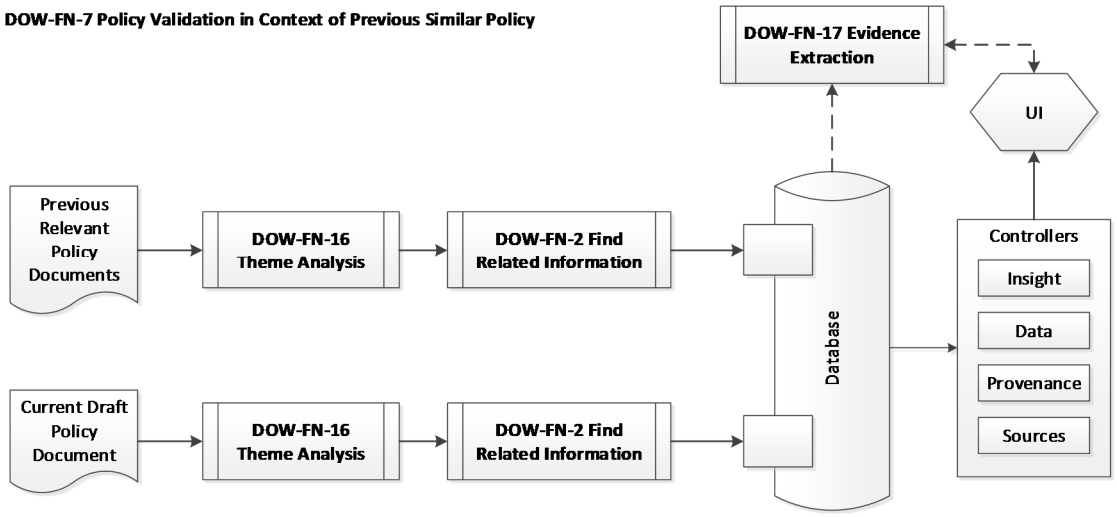


Figure 24: Policy Validation in Context of Previous Similar Policy Sub-Architecture

Figure 24 shows the sub-architecture for policy validation in the context of a previous similar policy. As can be seen, it is a case of finding related information for the current policy, and also previous policies. All data sets resulting will be stored in the database and can be compared. It may be that building an information model containing information derived from each policy document is a method for comparison. This is shown as optional in the figure, as it is up to the user how they want to use the tools to compare.

4.3.10 DOW-FN-5 Policy Model Construction

This scenario covers creation of policy models. A policy model is needed as input to the WP6 simulator and simulation enables the user to evaluate different policy options.

DOW-Function	Description
DOW Function ID	DOW-FN-5
Name	Policy Model Construction
Purpose	Create a policy model that can be used for evaluating different policy options.
Input	Information model, related information
Input DOW-Function(s)	DOW-FN-2 Find Related Information
Output	Policy model
Consumer DOW-Function(s)	Needed for DOW-FN-6 Simulation of Policy Option Impact
Operation Sequence	1. User has previously created an information model using related information discovered from their search terms or policy analysis. 2. User selects from elements and relationships in the information model for inclusion in the model. 3. User can run simulations on draft models and adjust models as necessary.

	4. Model and history of user actions to create the model are stored in the database. 5. User can view model, source data and provenance information stored in DB

Table 21: Policy Model Construction

F-ID	Description	Pri-ority	How to Address
F-9b	Show history of what user has done with data inside the toolkit	22	This will be supported by the UI and the infrastructure by showing a graph of data the user has collected or the outputs of analysis, along with the history path of how that data was generated or collected, showing e.g. which analysis tool was used, what its settings were, which data source was used, what the search terms were, etc.
F-13a	Transparency in the creation of the policy models	6	This will be achieved in a number of ways: by providing tools to help a user build a model for the policy problem, by structuring the problem in terms of the key variables and parameters identified by the user; allowing the user to define and modify links, policy objectives in terms of targeted changes in outcome indicators; and providing documentation of the methodology used for the simulation tool.
F-16a	The policy maker should be able to examine a policy model	6	The policy model is intended to be represented as an open graph of nodes and links, so the policy maker can examine its interior.
F-16b	The policy maker should be able to add new factors to policy model	6	The policy model is intended to be represented as an open graph of nodes and links, so the policy maker can examine its interior and identify any gaps. The other tools (e.g. the searches and analyses) are intended to help the policy maker find more related information, so these tools can be used to add nodes and links to the policy model graph.

Table 22: Relevant Functional Requirements for Policy Model Construction

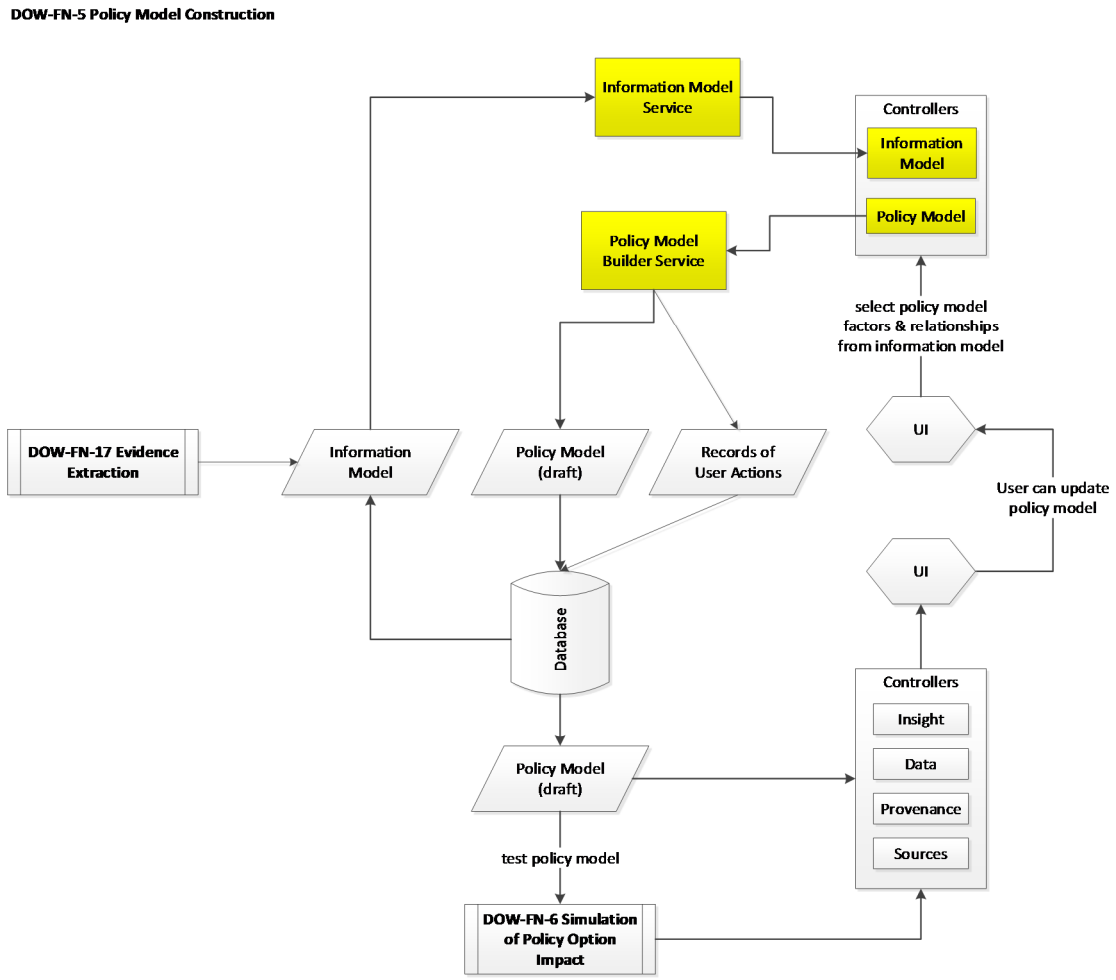


Figure 25: Policy Model Construction Sub-Architecture

Figure 25 shows the policy model creation. The user begins by looking at the information model and selecting some elements they want in the policy model. One of the likely criteria for selection of policy model candidates is a causal relationship between two elements, indicating that one element has some effect on another – this is a key property of the policy models used in WP6. Once the user has selected some elements, they can copy them into the policy model. The user may build up the policy model using elements from existing information models, or go back and delve further by gathering mode related information. The user can add elements to the information and policy models at any time, and in both, a history of the user's actions in creating and updating them is recorded, so the user can see what they did. The model may be tested by simulation (using DOW-FN-6, discussed next) and the results of the simulator may be examined, so that a user can make any updates necessary.

4.3.11 DOW-FN-6 Simulation of Policy Option Impact

This scenario covers the simulation of policy models to examine the impact of a particular policy option on dependent variables in the policy model.

DOW-Function	Description
DOW Function ID	DOW-FN-6
Name	Simulation of Policy Option Impact
Purpose	Simulation of the impact of a policy option using the simulator of U Stockholm and the policy model created by DOW-FN-5 as input.
Input	Policy model, simulation parameters, model parameter values
Input DOW-Function(s)	DOW-FN-5 Policy Model Construction
Output	Values for impacted variables in model
Consumer DOW-Function(s)	Needed for DOW-FN-14 Evaluation of Multiple Policy Options
Operation Sequence	1. User selects a policy model and chooses some input parameter values 2. User runs simulator 3. Results (values for impacted variables) are stored in database 4. User can view results and provenance information stored in DB

Table 23: Simulation of Policy Option Impact

F-ID	Description	Pri-ority	How to Address
F-13b	Transparency in the operation of the policy models	6	The simulation can be re-run a number of times, to enable the user to test different scenarios, and by adjustment of one parameter at a time, will enable them to determine the effect of that parameter.

Table 24: Relevant Functional Requirements for Simulation of Policy Option Impact

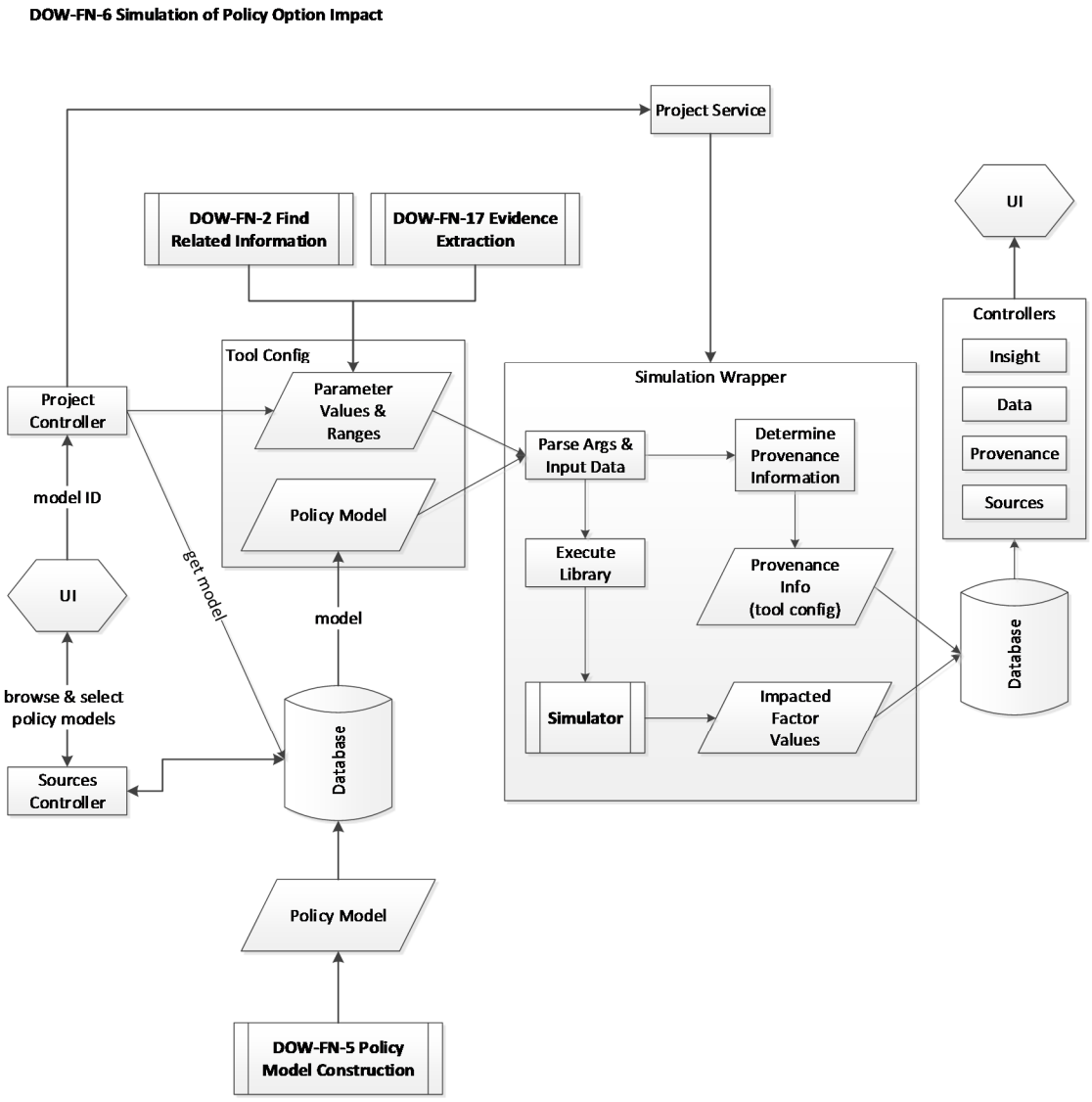


Figure 26: Simulation of Policy Option Impact Sub-Architecture

Figure 26 shows running of the simulator. The input (the policy model) is stored in the database and the user can select values and ranges for the model's input parameters. The simulator is wrapped in a similar fashion to other tool items for Sense4us, and the wrapper parses inputs, runs the simulation and stores the output and provenance information.

4.3.12 DOW-FN-14 Evaluation of Multiple Policy Options

This scenario covers extraction of keywords from source texts (documents or tweets are currently supported, but other types may be supported in the future).

DOW-Function	Description
DOW Function ID	DOW-FN-14
Name	Evaluation of Multiple Policy Options

Purpose	Multiple options for policies can be simulated by either choosing different input conditions for the same model, or different models.
Input	Multiple sets of: policy model, simulation parameters, model parameter values
Input DOW-Function(s)	DOW-FN-5 Policy Model Construction
Output	Keywords summarising input
Consumer DOW-Function(s)	Useful for DOW-FN-15 Generation of Search Terms
Operation Sequence	1. User select either a single policy model or multiple models 2. For each model, one or more sets of parameter values is specified. 3. User runs simulator for each input set (model + parameters) 4. Results (values for impacted variables) for each run are stored in database 5. User can view results and provenance information stored in DB

Table 25: Evaluation of Multiple Policy Options

F-ID	Description	Pri-ority	How to Address
F-13b	Transparency in the operation of the policy models	6	The simulation can be re-run a number of times, to enable the user to test different scenarios, and by adjustment of one parameter at a time, will enable them to determine the effect of that parameter.

Table 26: Relevant Functional Requirements for Evaluation of Multiple Policy Options

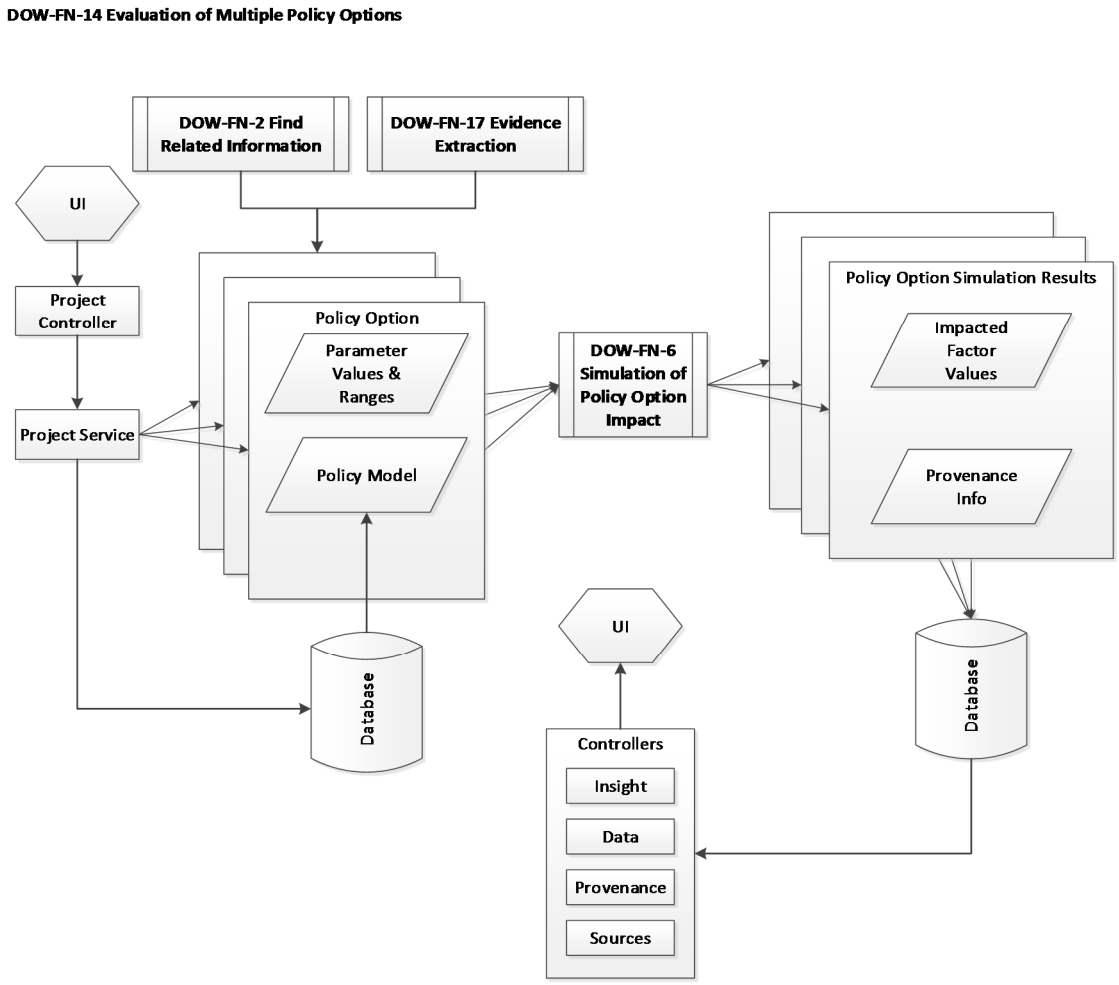


Figure 27: Evaluation of Multiple Policy Options Sub-Architecture

Figure 27 shows the sub-architecture for the evaluation of multiple policy options. As can be seen, it is the managed execution of DOW-FN-6 with different input models and parameters, which cover the testing space the user wants.

4.4 Requirements Addressed

This section describes which requirements have been addressed by DOW function scenario in the form of a matrix (Table 27). The functional requirements are listed as rows and the DOW function scenarios are columns. Where a DOW function scenario addresses a requirement, this is marked as a '1' in the intersection between them. The rightmost column is a total of the number of DOW function scenarios that address the requirement. The matrix highlights the requirements in red that have not been addressed by any scenarios (where the number of scenarios is zero), and these are discussed afterwards.

Functional Requirement	DOW-FN-01 Intelligent Search of Heterogeneous Sources	DOW-FN-02 Find Related Information	DOW-FN-05 Policy Model Construction	DOW-FN-06 Simulation of Policy Option Impact	DOW-FN-07 Policy Validation in Context of Previous Similar Policy	DOW-FN-08 Social Media Search	DOW-FN-09 Social Media Discussion Analysis	DOW-FN-13 Provenance Information Extraction and Display	DOW-FN-14 Evaluation of Multiple Policy Options	DOW-FN-15 Generation of Search Terms	DOW-FN-16 Theme Analysis	DOW-FN-17 Evidence Extraction	Number of Scenarios
F-01a Search different data formats	1					1							**2**
F-03b Enable the user to add specific websites as search targets	1					1							**2**
F-05a Enable opinions from social media to be gathered		1				1	1						**3**
F-05b Enable opinions from other forums to be gathered							1						**1**
F-06a Enable searching multiple data sources with the same query and mouse click from the user	1	1											**2**
F-09b Show history of what user has done with data inside the toolkit			1			1		1		1	1	1	**6**
F-10c Enable user to indicate trusted data sources								1					**1**
F-11a Enable the user to sort data by different criteria													**0**
F-13a Transparency in the creation of the policy models			1									1	**2**
F-13b Transparency in the operation of the policy models				1					1				**2**
F-14a Enable the searching for scientific information relevant to the policy subjects		1											**1**
F-15a Enable the collection of sentiments about the policy subjects	1	1				1	1						**4**
F-16a The policy maker should be able to examine a policy model			1									1	**2**
F-16b The policy maker should be able to add new factors to policy model			1									1	**2**
F-19a Discover any publicly available characteristic information about the social media participants who are expressing the opinions						1	1						**2**
F-20a Enable the user to specify different time ranges for when the data was (or is) relevant in a search	1					1							**2**
F-21b storage and retrieval of data within a user's account						1							**1**
F-21c Record workflows within a user's account								1					**1**
F-24a Ability to use data from the European Commission	1												**1**
F-24b Ability to use data from the European Parliament	1												**1**
F-25a (The system should) Augment search terms the user has specified		1								1			**2**
F-25b (The system should) perform automatic additional and related searches to the ones the user is running	1	1				1							**3**
F-26a Support multiple languages in searches	1					1							**2**
F-27a Provide summaries of raw data sets											1		**1**

Functional Requirement	DOW-FN-01 Intelligent Search of Heterogeneous Sources	DOW-FN-02 Find Related Information	DOW-FN-05 Policy Model Construction	DOW-FN-06 Simulation of Policy Option Impact	DOW-FN-07 Policy Validation in Context of Previous Similar Policy	DOW-FN-08 Social Media Search	DOW-FN-09 Social Media Discussion Analysis	DOW-FN-13 Provenance Information Extraction and Display	DOW-FN-14 Evaluation of Multiple Policy Options	DOW-FN-15 Generation of Search Terms	DOW-FN-16 Theme Analysis	DOW-FN-17 Evidence Extraction	Number of Scenarios
F-28a Visualisation of data sets													0
F-29a Highlight any data from social media													0
F-30b Filter by data origin													0
F-31a Determine if two data sets are comparable													0
F-31b Determination of how multiple data sets can be compared													0
F-32a Enable searches for data from countries outside the user's own						1							1
F-34a Present available metadata about a search result to the user						1		1					2
F-35a Enable the user to select which metadata they frequently want to see in their search result summaries													0
F-36a Locate discussions on social media that are related to the policy issue in question	1	1				1	1						4
F-37a Enable the user to specify a date that determines the earliest creation point of data in search results	1												1
F-38a In the event of multiple versions of a data set being available, highlight the most recent													0
F-40a Ability for the policy maker user to customise the tool to their preferences													0
F-41a The tool's UI should support German as a language													0
F-41b The analysis components should support German as a language		1				1	1			1	1		5

Table 27: Summary of Functional Requirements Addressed by DOW Functions

The following table is a cut-down copy of Table 3 (how to address functional requirements) to include only the requirements that are not addressed by any scenario.

F-ID	Description	Pri-ority	How to Address	Group
F-11a	Enable the user to sort data by different criteria	3	Results should be able to be sorted, and the criteria is likely to be many and varied. Results will be sort-able where possible, and this depends on the criteria. It is easy to sort numerically, alphabetically or chronologically, but other types of criteria will be examined on a case by case basis to determine whether it is possible. Where it is not possible to sort, the UI will indicate this	User Interface - Sorting

F-ID	Description	Pri-ority	How to Address	Group
			by greying out any "sort" buttons or tools.	
F-28a	Visualisation of data sets	3	This requirement is very general, as each data set is different and will have different visualisation criteria. In addition, there are many different visualisation types, and some are more suited to some data sets than others. Also, different users prefer to see things in different ways. Each data type encountered will be examined on a case by case basis to determine if it can be visualised, and how it can be visualised. This is potentially a vast task and needs to be made tractable, so it would be a good idea to evolve a standard set of visualisations, based on what is popular with end users, what is possible given the tools available, and what is compatible with the data.	Visualisation
F-29a	Highlight any data from social media	3	This is simply a case of showing the data source to the user and pointing out that it is social media.	User Interface - Filtering
F-30b	Filter by data origin	3	The user will be shown the data sources from which data came, and will be able to select those that should be shown in results.	User Interface - Filtering
F-31a	Determine if two data sets are comparable	3	It is expected that if this requirement is addressed at all, it will be addressed minimally and under tightly constrained conditions. Whether it is possible to address this requirement depends strongly on the data being compared and its compatibility. It is unlikely that automatic detection of compatible data sets will be possible, due to the multitudes of different data formats and semantics that may be encountered. Having said this, some fields may be possible to compare ("date" for example).	Visualisation

F-ID	Description	Pri-ority	How to Address	Group
F-31b	Determination of how multiple data sets can be compared	3	It is expected that if this requirement is addressed at all, it will be addressed minimally and under tightly constrained conditions. Whether it is possible to address this requirement depends strongly on the data being compared and its compatibility. It is unlikely that automatic detection of compatible data sets will be possible, due to the multitudes of different data formats and semantics that may be encountered. Having said this, some fields may be possible to compare ("date" for example).	Visualisation
F-35a	Enable the user to select which metadata they frequently want to see in their search result summaries	2	The user profile will contain a section for favourite metadata, and this allows the user to select metadata that will be presented in search results.	User Interface – Sorting
F-38a	In the event of multiple versions of a data set being available, highlight the most recent	1	This can be addressed by ranking the same dataset by creation date, so that the most recent is at the top.	User Interface – Language Support
F-40a	Ability for the policy maker user to customise the tool to their preferences	1	User profiles with customisation functions will be provided. There are many other functional requirement responses that contain customisation and user profile aspects (e.g. favourite filters and trusted source lists), and these will be gathered together to provide the specification for the user profile section of the Sense4us toolkit.	User Accounts & Customisation
F-41a	The tool's UI should support German as a language	3	The UI will be developed so that it may be skinned with different languages and a translation file supplied for German.	User Accounts & Customisation

Table 28: Functional Requirements Not Addressed by DOW Function Scenarios

The table is shaded because the requirements fall into three main groups:

The green group are all concerned with aspects of the user interface. The requirements for the UI are summarised as follows:

- Sorting
    - F-11a Enable the user to sort data by different criteria

    - F-38a In the event of multiple versions of a data set being available, highlight the most recent
- Filtering
    - F-29a Highlight any data from social media
    - F-30b Filter by data origin
- Language support
    - F-41a The tool's UI should support German as a language

It is not necessary to add a new sub-architecture, because these requirements are functions needed by one component (the UI), so it is enough to note that the UI should address these requirements.

The orange group is concerned with visualisation and representation of data. The requirements for this are summarised as follows:

- F-28a Visualisation of data sets
- F-31a Determine if two data sets are comparable
- F-31b Determination of how multiple data sets can be compared

These requirements are general in nature, and as the discussion in the table reflects, each data type being represented will need a new approach, unless it is possible to determine general patterns that cross data types. Visualisation will be supported on a tool-output basis – that is the output of different tools in Sense4us will have associated visualisations in the UI. Whether it is possible to compare two data sets strongly depends on the data sets themselves, for example it may be possible to compare some data set types, but others may be impossible. For the time being, this requirement will be investigated to determine if it is firstly possible, secondly for which data types, and thirdly if the effort is justified (it should be acknowledged that the requirement is one of the lower priority ones – the actual priority score is rather low).

The blue group is concerned with user accounts and customisation of the toolkit to the user's preferences:

- F-35a Enable the user to select which metadata they frequently want to see in their search result summaries
- F-40a Ability for the policy maker user to customise the tool to their preferences

As with the UI requirements above, these requirements are restricted to one component, the User Service, so a new sub-architecture is not required, and it is enough to note that the User Service should address these requirements.

4.5 Overall Architecture model

The overall architecture model has been developed by overlaying the sub-architectures corresponding to the individual scenarios and removing any redundant (repeated) components. The resulting model is shown in Figure 28.

The major observation is that the architecture is similar to the initial proposal as described in Figure 6. This is to be expected since the overall system design is based on the pattern

described in Figure 6, and the pattern incorporates the principles discussed in Section 3. In addition, the pattern has been implemented in the demonstrator of D7.1.

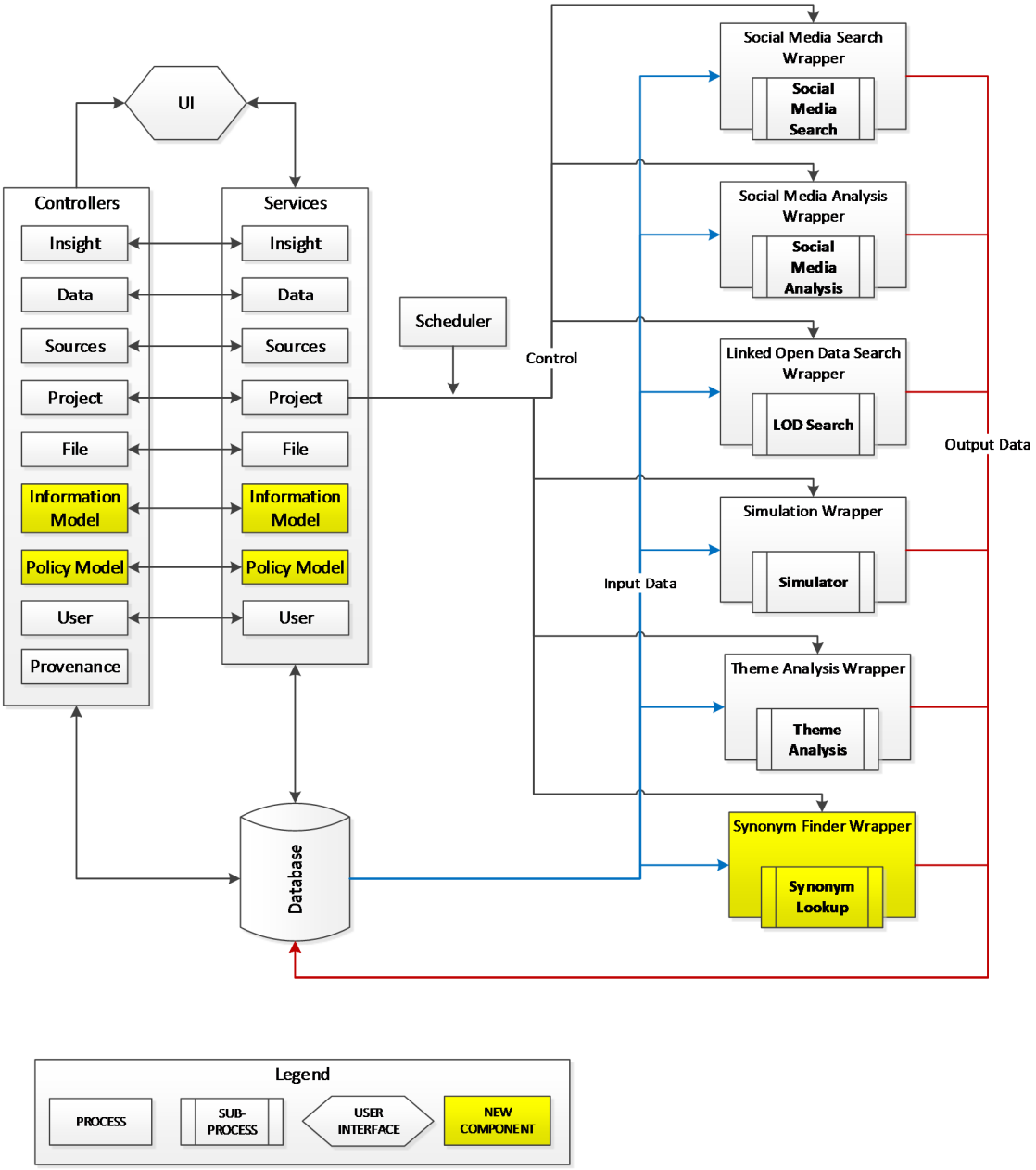


Figure 28: Overall system architecture

5 Summary

In this deliverable we have presented the architecture of the Sense4Us system. In addition, we have presented a process to determine the architecture that takes into account a set of differing inputs:

- functional requirements derived from end user requirements;
- the original concept of the project and the project's initial architectural principles; and
- the initial demonstration prototype.

The architecture has been derived based on D3.1 (Functional Specification), which itself is in turn based on D2.1 (user requirements). The Description of Work (DoW) is also a key input to the process as it determines the project's overall concept. The project's concept includes the project's research work conducted by partners in WPs 4, 5 and 6, and so this research was also taken into account. The architecture also has background work as input in the form of a basic infrastructure that has been proved in previous project and has been developed further within Sense4us, to produce the initial demonstrator forming D7.1.

The architecture itself has been determined through the definition of use case scenarios that represent the core functionality of the project and how it may be used by end users. Functional requirements were analysed to remove redundancy and to determine the ones that needed addressing, and requirements were allocated to the scenarios that could address them. For each scenarios, a "sub-architecture" was created, which shows the components from the initial demonstrator and the components from the research workpackages. Each scenario's sub-architecture shows how these components may be used to deliver the functionality described by the DOW. This process highlighted where any components were missing, and these have been described. The final step in the architectural determination process was to combine all the scenario sub-architectures into one master architecture, which is the outcome of this deliverable. This is the first version of the system architecture and will be updated in month 30, in D3.4.