COBWEB PROJECT
Deliverable Report

Grant Agreement number: 308513
Project acronym: COBWEB
Project title: Citizen Observatory Web
Funding Scheme: FP7 ENV.2012.6.5-1
Project website address: www.cobwebproject.eu
Deliverable: D4.3 “Research findings and specification for the conflation of the authoritative and crowdsourced data sources”
Projected delivery date: M18
Actual delivery date: M19
Lead partner: Technische Universität Dresden (TUD)
Coordinating Editor: Stefan Wiemann (TUD)
Reviewer: Michael O’Grady (NUID UCD)

This project has received funding from the European Union’s Seventh Programme for research, technological development and demonstration under grant agreement No 308513
# Table of Contents

1. Conventions ........................................................................................................... 1
   - Abbreviations .................................................................................................... 1
   - Notations ........................................................................................................... 1
   - Definitions ......................................................................................................... 1

2. Executive summary .................................................................................................. 3

3. Conflation sub-system for COBWEB ................................................................. 4
   1.1. State-of-the-art in conflation ........................................................................... 4
   1.2. Conflation implementations and algorithms ................................................. 9
   1.3. COBWEB conflation targets ......................................................................... 12
   3.1. Requirements from the COBWEB case studies ........................................... 13
   3.2. Available data sources for conflation ............................................................ 18

4. Conflation use cases ............................................................................................... 20
   3.4. Provide Reference Data ............................................................................... 20
   4.1. Collect and Upload Data .............................................................................. 22
   4.2. Conflate Data ............................................................................................... 23
   4.3. Interlink Data ............................................................................................... 24
   4.4. Access Reference Data ................................................................................ 25
   4.5. Access Conflated Data ................................................................................ 26
   4.6. Visualise Conflated Data .............................................................................. 26
   4.7. Conflation architecture ................................................................................ 29
   5.1. Service components ..................................................................................... 29
   5.2. Workflow ...................................................................................................... 30
   5.3. Quality control .............................................................................................. 33

5. Conclusion .............................................................................................................. 35

6. References ............................................................................................................... 36

7. Appendix 1: Document History ........................................................................... 41
Tables
Table 1: Summary of conflation service approaches for mobile devices ........................................17
Table 2: Possible data sources for conflation ..................................................................................18

Figures
Figure 1: Classification of conflation operations ............................................................................. 5
Figure 2: Decomposition of the conflation process ........................................................................... 6
Figure 3: Basic principles of the conflation targets ............................................................................ 12
Figure 4: Overview on conflation use cases ...................................................................................... 20
Figure 5: Conflation core components for COBWEB ....................................................................... 29
Figure 6: Basic workflow for Linked Data based conflation ............................................................... 31
Figure 7: Interaction between conflation and quality assessment ..................................................... 34

List of contributors

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stefan Wiemann</td>
<td>TUD</td>
</tr>
<tr>
<td>Lars Bernard</td>
<td>TUD</td>
</tr>
<tr>
<td>Pierre Karrasch</td>
<td>TUD</td>
</tr>
</tbody>
</table>
1. Conventions

Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COBWEB</td>
<td>Citizen Observatory Web</td>
</tr>
<tr>
<td>INSPIRE</td>
<td>Infrastructure for Spatial Information in Europe</td>
</tr>
<tr>
<td>GML</td>
<td>Geographic Markup Language</td>
</tr>
<tr>
<td>JSON</td>
<td>JavaScript Object Notation</td>
</tr>
<tr>
<td>OGC</td>
<td>Open Geospatial Consortium</td>
</tr>
<tr>
<td>RDF</td>
<td>Resource Description Framework</td>
</tr>
<tr>
<td>SOS</td>
<td>Sensor Observation Service</td>
</tr>
<tr>
<td>SPARQL</td>
<td>SPARQL Protocol and RDF Query Language</td>
</tr>
<tr>
<td>UML</td>
<td>Unified Modeling Language</td>
</tr>
<tr>
<td>WMS</td>
<td>Web Map Service</td>
</tr>
<tr>
<td>WFS</td>
<td>Web Feature Service</td>
</tr>
<tr>
<td>WPS</td>
<td>Web Processing Service</td>
</tr>
<tr>
<td>XML</td>
<td>Extensible Markup Language</td>
</tr>
</tbody>
</table>

1.1. Conventions

1.1.1. Abbreviations

1.1.2. Notations

UML (Activity, Use Case): The UML 2 standard is used to describe activity workflows and use case structures for conflation. A detailed description of the standard and graphical notations is provided by the Object Management Group1.

1.1.3. Definitions

Conflation: The term conflation was first associated with spatial data in a study by the USGS (United States Geological Survey) and the Bureau of Census in the 1980s to consolidate digital maps based on feature matching and positional re-alignment (Saalfeld 1988). Based on the definition of Yuan and Tao (1999), the term conflation hereinafter describes the combination of spatial data from multiple sources to produce a combined view that contains the most valuable data from the inputs. Valuable hereby

1 http://www.omg.org/spec/UML/2.4.1/Superstructure/PDF

This project has received funding from the European Union’s Seventh Programme for research, technological development and demonstration under grant agreement No 308513.
depends on the application context. This task is often also referred to as data fusion or data integration (Ruiz et al. 2011).

**Homologous Feature**: Since spatial data is an abstraction of the real world, different spatial concepts and data acquisition means can lead to a multitude of possible feature representations. If two features describe the same real world object, they are defined as homologous. In a part-of relation, only the overlap of two features is considered homologous.

**Linked Data**: The Linked Data paradigm represents a major building block for the Semantic Web. It is realised by (1) the use of Unified Resource Identifiers (URIs) to uniquely identify objects or concepts that can (2) be resolved to HTTP addresses providing (3) data in standardised formats that are (4) interlinked to other data sources on the web (Berners-Lee 2006).

**Matching**: The automated detection of related features, feature elements, schemas or ontological concepts is defined as matching. It is applied to relate objects within a dataset or between different datasets, and thereby provides information on their relationships.

**Schema Mapping**: A schema mapping defines relations between source and target schema elements. It implicitly expresses mapping rules that allow for the transformation of data on the syntactic, schematic and semantic level. According to the Technical Guidance for INSPIRE Schema Transformation Services (INSPIRE 2010b), the Rule Interchange Format (RIF) is proposed as most suitable for encoding schema mapping definitions within SDI.

**Sensor Fusion**: The term sensor fusion is primarily used within the remote sensing domain and describes the synthesis of multiple sensors in order to receive an enhanced view on the observed phenomenon or entity (OGC 2010a). Within this document, the term is aligned to the COBWEB project and describes the combination of spatial and temporally related sensor measurements on the same environmental phenomenon that can be used to enhance and validate crowdsourced data.
2. Executive summary

This deliverable summarises the research findings for the conflation of crowdsourced and authoritative data to be applied for the implementation of a COBWEB conflation sub-system. This sub-system shall help the users and scientists to compare, validate and integrate environmental observations with existing, primarily authoritative, reference data to support further analysis and decision making.

Conflation in general consists of a number of sub-processes, including data search and retrieval, data enhancement and harmonisation, similarity measurement, data matching, evaluation and resolving. For each of those steps, appropriate methods and tools need to be developed. Following the COBWEB functional architecture (D3.1), a service-based approach is required, offering conflation functionality via an open standardised interface. This can be achieved by using the OGC Web Processing Service standard.

The main conflation targets for COBWEB are data densification, enrichment and update. For each of those use cases, relations between crowdsourced and authoritative data need to be identified and stored. Therefore, the Linked Data paradigm is applied for the storage of feature links and associated metadata on provenance and quality. This enables a customised conflation based on specific user or case study requirements.

The presented architecture is aligned to the general COBWEB framework and incorporates external components, such as COBWEB data access services, the COBWEB registry as well as services offering functionality for quality assurance. It specifies components for the provision of conflation functionality, the visualisation of conflation results and a data repository for storing intermediate and final conflation results. Moreover, it is shown that conflation can be a sub-task of quality assurance and vice versa.
3. Conflation sub-system for COBWEB

The COBWEB project focuses on developing web-based conflation capabilities for linking crowdsourced environmental information with existing (for example authoritative) reference data to create value-added information from its combination. The aim is to achieve an integrated source of spatial information, which is independent of underlying data sources and method of data access. The resulting technology stack shall be used to support the case studies defined by WP2. In general, the interoperable, flexible and automated combination of spatial data sources on the web is considered as a major building block for the Semantic Geospatial Web as proposed by Egenhofer (2002).

Conflation can be used for a number of different purposes within the COBWEB system. In general, it needs to be applied, where information is required that cannot be obtained from one or several isolated datasets.

State-of-the-art in conflation

As shown in Figure 1, conflation operations can be classified based on a number of criteria (Yuan and Tao 1999, Schwinn and Schelp 2005, Ruiz et al. 2011). Those include the:

- operation frequency, distinguishing between unique or rare events (usually performed manually), periodic events (semi-automated) and real-time events (fully automated)

- semantic level, which can be the feature representation (geometric and thematic attributes), feature type (attribute definitions, schema structure) or ontology (conceptual definitions)

- data type of the input sources, a combination of either vector or raster based data structures,

- operation direction, which is typically vertical (overlay of data covering the same spatial extent), but can also be horizontal (edge matching) or temporal (same dataset, but with different timestamps)
Decisions concerning the appropriate process primarily depend on the application purpose. Here applications can be distinguished as focussing:

- detection of changes, discrepancies and errors
- data update and enrichment
- full integration of multiple spatial datasets

In addition, the differentiation of processes can be based on supported input and output formats, average precision and recall rates of the matching strategy or certain quality of service elements, such as the computational performance.

To meet the requirements of an SDI-based and possibly real-time conflation of crowdsourced data with existing reference data, a service-based approach for conflation is required. A first step towards the design of services is the identification and definition of proper service interfaces and an appropriate granularity of the operations, which are offered by the services. Therefore, complex processes need to be decomposed into well-defined atomic processes in order to match the requirements of a Service Oriented Architecture (SOA, Erl 2008). An example on how service-based conflation can be implemented is given by Wiemann and Bernard (2010).

Spatial data conflation originally comprises the two steps of 1) measuring similarity to identify potential homologous features and 2) positional re-alignment of the considered geometries (Rosen 1985, Lupien 1987, Saalfeld 1988, Doytsher et al. 2001). These
core processes have been complemented by additional processes, such as preparatory data harmonisation (Uitermark 1999, Zhang 2007, Stankute 2012), quality and confidence measures (Cobb 1998, Walter 1999, Bloch 2001) and feature attribute transfers (Yuan and Tao 1998, Blasby 2003).

Based on the existing approaches, a subdivision of conflation into seven sub-processes is proposed (Figure 2). However, those sub-processes are not meant to be clearly separable, because a single process might fall into more than one of the suggested categories. Furthermore, the workflow might change, as components can be optional, iterated or concatenated in different ways. Nevertheless, the proposed concepts are meant to serve as an overarching framework for service-based conflation of spatial data.

**Figure 2: Decomposition of the conflation process**

**Data search and retrieval** is typically the first step in data conflation and is required to identify and get access to input datasets for conflation, for COBWEB in particular authoritative data and crowdsourced environmental observations. While data search mainly deals with metadata, the retrieval primarily addresses data provision means and encodings. To facilitate interoperability, open standards should be used for both search and retrieval of data. Within SDIs, this can be realised by using the OGC specifications on the Catalogue Service for the Web (CSW, OGC 2007) and OGC data service standards, such as the Web Feature Service (WFS, OGC 2010c), the Web Coverage Service (WCS, OGC 2010b) or Sensor Observation Service (SOS, OGC 2012). Spatio-temporal and thematic filters can be applied to reduce the network load and support further processing.

**Data enhancement** mainly addresses the quality characteristics of single input data. It comprises data curation, statistical analysis and structural improvements with the aim to achieve a certain level of quality and applicability for further processing. In the majority of existing applications, this step is considered as preparatory work or preprocessing. Within the COBWEB project, a certain quality level for the input data needs to be assured by the data provider (in case of authoritative data) or the COBWEB system (in case of crowdsourced data).

This project has received funding from the European Union’s Seventh Programme for research, technological development and demonstration under grant agreement No 308513
Data harmonisation is required to align datasets and addresses syntactic, structural as well as semantic differences (Mohammadi 2010). Examples are positional feature re-alignment (Lupien and Moreland 1987, Doytsher et al. 2001, Song 2006, López-Vázquez and Callejo 2013), time synchronisation (Uitermark 1999), feature extraction (Chen 2008) format and coordinate conversions (Stankute 2012) or model generalisation (Sester 1998, Uitermark 1999). Corresponding constraints, determining the scope for harmonisation, can be created using similarity measurements on the dataset level.

Similarity measurement investigates and quantifies the relationship between the inputs. Corresponding measures can be performed on the feature representation, the feature type and the domain ontology level. A collection of frequently used approaches are summarised by Veltkamp (2001), Rahm and Bernstein (2001), Ruiz et al. (2011) and Delgado (2013). The result of similarity measurements are usually stored within similarity sets (Samal et al. 2004) or matrices (Wong 1995, Rahimi 2002, Song 2011).

Data matching is closely linked to the similarity measurements and can accordingly be classified into feature-, schema- and ontology-based approaches. Matching on the feature level exploits spatial, temporal and thematic characteristics of feature representations. A number of corresponding application use cases are summarised by Ruiz et al. (2011). The matching of schemas and ontologies is based on terminological, syntactic, structural and conceptual characteristics. A review of corresponding approaches is presented by Rahm and Bernstein (2001), Shvaiko (2005) and Bleiholder and Naumann (2008) for schema and Buccella (2009) and Delgado (2013) for ontology matching. In addition, there are multi-layered approaches, such as presented by Volz (2005) combining feature and schema matching or Du (2012) combining feature and ontology matching. The results are often stored internally, but can also be formalised using a custom XML schema (Volz 2005) or RDF representation (Volz 2009, Stadler 2012).

Evaluation of results is tightly coupled to the matching process and addresses feature, schema and semantic conflict or change detection as well as the identification of homologous features, mostly related to certain confidence measures (Cobb 1998, Bloch 2001, Waltz 2008). In addition to the underlying similarity measurements, the evaluation of matches might also depend on the reliability of input sources or common knowledge (Cholvy 2007). The evaluation results are usually attached to the identified input matches (see previous step) to assist in the final resolving process (see next step).

Resolving includes data merging, feature and attribute transfers and eventually post-processing steps such as coordinate and format conversions. Furthermore, strategies for conflict resolution can be applied for data de-duplication or
handling missing data (Cobb et al. 1998, Bloch et al. 2001, Bleiholder and Naumann 2008). Finally, the resolving process sends the conflated result to the consumer, ideally with associated metadata on the conflated process, input data and quality measures.

The following fictitious and simplified example shall illustrate the workflow and therefore describes the conflated of a number of species observations taken by volunteers in the field with an official species distribution map provided by a local authority. The data is assumed to be structured as follows:

(1) Species observations: points with GPS coordinates (WGS84); attribute ‘name’ contains name of the species; attribute ‘timestamp’ contains timestamp of the observation; obtained from multiple users

(2) Species distribution map: areal features; ETRS-LAEA projection; attribute ‘species’ contains species name; attribute ‘sightings’ contains number of sighting for the particular species in the area; data is provided via OGC WFS interface – one layer per species

In accordance with the sub-processes identified above, the conflated of both datasets could be performed as follows:

1. The conflation process is triggered by sending (1) to the conflation service.
2. Based on the ‘name’ attribute of (1), corresponding WFS layers are identified and requested; (2) is retrieved using a WFS GetFeature request using a spatial filter derived from the extent of (1). [data search and retrieval]
3. Quality control is performed for (1) to remove erroneous observations (for example name cannot be resolved to a species). [data enhancement]
4. A coordinate transformation is performed to transform (1) into ETRS-LAEA projection and ensure comparability with (2). [data harmonisation]
5. Distance measurements are carried out between the observations in (1) and between the features in (1) and (2). [similarity measurements]
6. The features in (1) and (2) are matched. Between observations in (1) ‘sameAs’ relations are created, based on a threshold distance between coordinates and the ‘name’ attribute. Between (1) and (2) ‘belongTo’ relations are created, if an observation lies within a species distribution area or at least within a specified threshold distance. [data matching]
7. Observations from (1) that have no relation to any area from (2) are identified. Furthermore, a confidence measurement, based on position and timestamp, is
performed for the ‘sameAs’ relationships between the observations in (1) to assess the probability of a relation. [evaluation]

Observations from (1) are neglected, if they do not have a relation to any area from (2) and no verification by another sighting within a ‘sameAs’ relationship. Multiple observations connected in a ‘sameAs’ relationship, but outside of an area from (2) are used to trigger a verification request to the local authority in order to verify the sighting and update (2) accordingly. Observations from (1) with a ‘belongsTo’ relationship to an area in (2) are counted and added to the ‘sightings’ attribute in (2) [resolving]

Conflation implementations and algorithms

Prior to building a conflation system, existing implementations and algorithms should be considered to avoid reimplementation efforts and known pitfalls. Previous work, in particular on the conflation of crowdsourced and authoritative data, is summarised in this section.

3.2. Previous work – literature review

Since the OpenStreetMap (OSM) project is one of the most prominent and supported crowdsourcing projects for voluntarily collecting spatial information, it is frequently used for comparison and conflation with authoritative reference data. Haklay (2010) and Girres and Touya (2010) assess the quality of OSM by statistical analysis, both with focus on the accuracy (geometric, attribute, semantic & temporal), completeness, logical consistency, lineage and use of OSM. Both studies determine the quality based on a number of geometry and thematic attribute measurements. Whereas the analysis of Haklay (2010) is primarily based on global feature statistics, Girres and Touya (2010) analyse the quality on the feature level. However, the matching of homologous objects is carried out manually and thus, only covers a small fraction of the OSM database. As a result, both emphasise the need for further research on quality assurance for crowdsourced data to make it comparable and competitive to existing data sources and solve the inherent conflict between data quality and amateur contributions. The same conclusion is drawn by Fairbairn and Al-Bakri (2013), who compared OSM data to two other official datasets. Especially for large scale maps, an integration of VGI and authoritative sources is considered to be a very sensitive problem. However, Pourabdollah et al. (2013) show that both sides can very well benefit from each other by highlighting the similarities and differences.

Du et al. (2012) describe a matching of road networks using its topological structure. Therefore, OSM data is compared and conflated with road data from the Ordnance Survey using a number of geometric and label matching algorithms. Basic consistency checks and conflict resolution strategies are applied to combine the input sources and generate the result. Yang et al. (2013) and Yang et al. (2014) implement a matching approach for crowdsourced data (OSM road network & graph representation of POIs)
with a professional road reference network using a probabilistic relaxation approach. Initial matches are calculated using geometric similarity measures (for example direction, distance, length). Afterwards, matching candidates are iteratively refined (confirmed or refused) by calculating a custom probability coefficient. The final matching pairs are selected based on the maximum probability and filters for multiple or null matches.

Wiemann and Bernard (2010) describe a service based approach for spatial data conflation within SDI. Therefore, a number of processing services were implemented, using the WPS interface, to compare data from OpenStreetMap with reference data from ATKIS (German Authoritative Topographic Cartographic Information System). The matching is performed using a number of geometric and attribute similarity measurements and encoded in a custom XML format. Based on the matches, attribute transfers are performed to generate added value for the reference data. Furthermore, basic confidence information is associated to the matches and stored with the final results.

As a result, it can be stated that a lot of attention has been given to the integration of crowdsourced information with authoritative sources. However, the current focus is primarily on road networks, in particular including data from the OpenStreetMap project, and associated quality issues. In the COBWEB context this needs to be adapted and extended, to enable the integration and validation of crowdsourced environmental observations with authoritative reference data.

3.2.2. Commonly used operations
The conflation of spatial data mainly builds upon similarity measurements and the subsequent identification of feature relationships. As mentioned before, both can be conducted on the feature representation, schema or ontology level. Commonly applied and adapted algorithms are summarised below:

- Spatial distance measurements:
  - Geographic distances are applied between points, to identify nearest neighbours. Since exact distances on the earth’s surface are difficult to calculate, the geographic distance is usually based on a flat surface (for small distances) or sphere. However, if appropriate, ellipsoidal deviations or elevations can be considered.
  - The average distance between two polylines (P) describes the ratio between the surrounded area (A) between and the length (L) of the polylines.
    
    \[ dA(P_1P_2) = A_{P_1P_2} / (L_{P_1} + L_{P_2} / 2) \]
The Hausdorff distance can be applied to calculate the maximum minimal distance between two point clouds \( (C) \). Although it does not provide sufficient information on the actual similarity of a shape, it is a good estimate for positional equality.

\[
dH(C_1C_2) = \max \left\{ \max_{x \in C_1} \left\{ \min_{y \in C_2} d(x,y) \right\}, \max_{x \in C_2} \left\{ \min_{y \in C_1} d(x,y) \right\} \right\}
\]

The Fréchet distance is a parameterised distance measurement between two polylines \( (P) \). It usually provides better information on the equality of shapes, but demands higher implementation efforts in comparison to the Hausdorff distance.

\[
dF(P_1P_2) = \min_{\alpha,\beta} \left\{ \max_{t \in [0,1]} \left\{ d(P_1(\alpha(t)), P_2(\beta(t))) \right\} \right\}
\]

The surface distance between two polygons \( (G) \) can be calculated as the ratio between the intersection and union areas \( (A) \).

\[
dS(G_1G_2) = 1 - \left( \frac{A(G_1 \cap G_2)}{A(G_1 \cup G_2)} \right)
\]

- Descriptive statistics that can be applied to assess the comparability of datasets or feature collections. Those can include the number of features and vertices, the average length and linearity of features or a comparison of timestamps associated with features.

- Topological relationships between features, based on a well-defined classification, such as the Egenhofer 9-intersection metric (Egenhofer and Franzosa 1991).

- The Spider Function (Rosen and Saalfeld 1985) characterises nodes within a linear network by the number of outgoing edges and their corresponding angles. It results in a characteristic pattern for each node, which can be compared to each other.

- The Iterative Closest Point (ICP) algorithm is frequently used in the remote sensing domain and can be used to support the matching of point clouds. Therefore, an affine transformation is iteratively applied to minimise the mean square error between the points and finally assign suitable matches.

- Buffer Growing (Walter and Fritsch 1999) takes into account the different segmentation of linear features and is thus well suited for matching linear networks. Subsequent to the identification of an initial match, a buffer is applied to find additional matches in the spatial surrounding, until a certain termination condition is reached.
String distance measurements:

- The *Damerau-Levenshtein* distance is used to calculate the similarity between character strings by counting the operations required to transform one string into the other. Those operations include the insertion, deletion, substitution and transposition of characters.

- The *Jaro-Winkler* distance calculates the similarity of two strings by counting the number of matching characters within a pre-defined search window.

**COBWEB conflation targets**

Within COBWEB, three specific targets for conflation are identified: (1) the densification, (2) the enrichment and (3) the update of spatial data *(Figure 3)*.

![Figure 3: Basic principles of the conflation targets](image)

**3.3. Densification of spatial data**

The densification of data is used to refine a response to a user request on describing a specific environmental phenomenon. At first, existing reference data is searched which matches the information request. If the data being found is not yet sufficient in terms of its spatio-temporal or thematic resolution, crowdsourced data is searched and, if applicable, combined with the previous result to enhance the response accordingly. Finally, the combined data set is provided to the user and includes lineage and quality information from the input data streams.

**3.3.2. Enrichment of spatial information**

Data enrichment mainly targets adding thematic information on top of an existing dataset. Accordingly, a result of a data request includes additional information on related datasets or observations. Those may help to validate the data or support the identification of additional data sources that might be relevant to the user as well.
Furthermore, existing environmental data can be related to external informational sources on the web to facilitate its usability.

3.3.3. Spatial data update
Spatial data updates can be applied to enhance the information provided by existing datasets. Since data providers usually have limited resources to regularly update information, crowdsourced information can give hints on where data needs to get updated. Therefore crowdsourced information can systematically and regularly be compared to existing datasets, providing information on missing, incomplete, outdated or erroneous parts, mainly based on change detection methods. The data provider can directly identify those parts and specifically update the affected data sets.

Requirements from the COBWEB case studies
A survey from the COBWEB case studies revealed a number of functional and non-functional requirements that have to be considered by the conflation sub-system to be implemented. Since the case studies and the related demonstrators are being continuously developed and updated during the course of the project, the specific requirements may change accordingly. Thus, the requirements described here are considered generic and not bound to any specific use case or technical specification.

3.4. Functional requirements

Provide functionality for conflation: As specified in the COBWEB Functional Architecture (D3.1), conflation functionality needs to be provided as open standardised web services. Since COBWEB will mainly build upon OGC standards, the use of the OGC WPS (OGC 2007) is envisaged. The WPS provides well-defined process descriptions and functionality within an SDI and thus, facilitates a loose coupling to other COBWEB system components, in particular for quality assurance. The conflation services will be provided on different composition layers, comprising low-level conflation services (as described in chapter 3.2.2), high-level services (complex conflation functionality or a composition of low-level services) and use case services (composition of low and high-level services bound to a specific use case).

Provide access point(s) for conflated data: The results of a conflation need to be provided for further use or visualisation purposes. Results can be obtained directly from the WPS, encoded in a standardised format (for example GML, RDF, GeoJSON), or stored in a data repository for later access using a data download or visualisation service.

Provide discovery mechanism for conflation services and conflated data: To publish, search and bind conflation services, a registry for WPS services is required. Furthermore, process descriptions need to be formalised to facilitate the automated search and the orchestration of the services. Conflated data can be published directly
with the processing services or, if stored in a COBWEB repository, as data download or visualisation services in accordance with the general COBWEB architecture.

*Provide capabilities to access data sources on the web for conflation:* Conflation services can be directly bound to input data or take input data as a process parameter. Therefore, the services need to support the spatial data encodings used within the COBWEB system. As specified in COBWEB D7.1 “Data Management Guidelines”, this especially applies to XML and JSON encodings, but may as well include the OGC GeoPackage (OGC 2014) standard.

### 3.4.2. Non-Functional Requirements

**Computational performance:** Depending on the application purpose and data volume, the services should deliver its results in a reasonable amount of time. This is especially true for real-time conflation use cases. However, a balance must be found between conflation quality and service response time.

**Quality information awareness:** All conflation services need to be capable of reading, tracking and, if applicable, processing quality information (compliant to ISO19115/19157) attached as metadata to the input sources. The conflation results should include quality information accordingly.

**Reliability:** To gain trust and confidence, the reliability of the provided conflation functionality need to be assessed. This can best be achieved by providing detailed and standard-compliant descriptions of the conflation processes and outputs compliant to ISO19119. Furthermore, a good service uptime and documentation can help to improve service reliability.

**Traceability:** All input data and processing steps should be documented in the metadata attached to the conflation results (data provenance and lineage). This allows for the reproducibility of all conflation steps, including access to original input data and intermediate results.

**Proper exception handling:** If an error occurs during process execution, a meaningful error report and log should be returned. This enables the user, developer or service administrator to identify the source of an error and, ideally, provide hints on how it can be solved.

### 3.4.3. Visualisation of conflated data

To implement a visualisation component on top of conflation processes, the following questions need to be addressed:

- What kind of conflation results can be visualised?
- By which means can conflation results be visualised?
How can conflation quality and lineage information be displayed?

The amount of data available for visualisation depends on the way it is provided and their access constraints. The citizen observations as the main input for conflation are part of the COBWEB system and can thus, directly be accessed for visualisation. However, access to observations made by different users might be restricted to some extent to assure privacy. However, if observations are compared to additional datasets, such as authoritative reference data or satellite imagery, access restrictions, limiting visualisation capabilities for the end user, may certainly apply. This is especially true for the results of a conflation process, which inherits the access restrictions from any of its sources. The following types of restriction may occur:

- **No restrictions** – information can be visualised on the client without any restriction,
- **No access to feature data** – information can be visualised as a map display, but the underlying feature data cannot be accessed,
- **Selective access to data** – parts of the data are allowed to be visualised; corresponding filters need to be implemented by the COBWEB system,
- **Generalised access to data** – information is allowed to be visualised on a certain scale; corresponding model generalisation needs to be implemented by the COBWEB system,
- **No access** – the data may not be visualised or accessed.

To access conflation results on a client, an agreed standard for data transfer and encoding should be applied. Depending on the use case and data restrictions, this can be either a data or map visualisation service. If applied in the field, sufficient internet connectivity and support for the applied communication standard is required on the mobile device. In general, the conflation results can be provided using one of the following formats:

- **Feature data** – spatial objects or coverage data that can directly be accessed and rendered by the client. Attributes and additional information on the features can be used for symbolisation and accessed by the user. However, the rendering of spatial data on the client side is relatively costly and may require additional dependencies.
- **Image data** – pre-rendered images served by a remote service that can directly be displayed by the client. Whether additional information is accessible to the user depends on the capabilities of the implemented service interface.
• **Descriptive data** – textual information on the conflation process, such as notifications on the validation of observations made by a user, which can be displayed in a text-field. Descriptive data may also include metadata for the conflation results and link to the conflation inputs or additional sources on the Web.

Visualisation capabilities on mobile devices are quite limited with respect to size, resolution and interaction. To facilitate the readability, displayed quality information should not be too complex and follow a simple colour schema indicating quality levels (for example red: poor quality, green: good quality, grey: unknown quality). Additional information, such as feature and processing lineage, should be visualised separately, and only on request, to not overload the display.

### 3.4.4. Conflation services on mobile devices

In COBWEB, the majority of environmental observations will be made in the field using mobile devices, such as sensor stations or smartphones. If conflation functionality is required at this place, corresponding access mechanisms have to be explored. Although focussing on conflation here, the same applies for other complex spatial data processing (for example quality control), which needs to be applied on the crowdsourced data.

For implementing conflation processes for mobile devices, there are basically 3 ways: (1) the use of a processing library offering conflation functionality directly on the mobile device, (2) the use of a processing service running directly on the mobile device and (3) the application of a remote processing service. A characterisation of the approaches is summarised in
This project has received funding from the European Union’s Seventh Programme for research, technological development and demonstration under grant agreement No 308513
Table 1: Summary of conflation service approaches for mobile devices

<table>
<thead>
<tr>
<th></th>
<th>Processing library</th>
<th>Local service</th>
<th>Remote service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network connection</td>
<td>Optional, required only for updates</td>
<td>Optional, required only for updates</td>
<td>mandatory</td>
</tr>
<tr>
<td>Extensibility</td>
<td>via software update</td>
<td>via transactional service protocol</td>
<td>via transactional service protocol</td>
</tr>
<tr>
<td>Reusability</td>
<td>on the mobile device</td>
<td>on the mobile device</td>
<td>on the web</td>
</tr>
<tr>
<td>Data transfer</td>
<td>minimal</td>
<td>minimal</td>
<td>network transfer</td>
</tr>
<tr>
<td>Access restrictions</td>
<td>limited to open data</td>
<td>limited to open data</td>
<td>Possible use of restricted data</td>
</tr>
<tr>
<td>Computational Performance</td>
<td>low</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td>Workflow management</td>
<td>tightly coupled</td>
<td>loosely coupled</td>
<td>loosely coupled</td>
</tr>
<tr>
<td>Dependencies</td>
<td>need to be stored locally</td>
<td>need to be stored locally</td>
<td>managed by the server</td>
</tr>
<tr>
<td>Maintenance</td>
<td>via software update</td>
<td>via software update</td>
<td>by service provider</td>
</tr>
</tbody>
</table>

Based on the mentioned pros and cons in
Table 1, the following conclusions are drawn for the COBWEB project:

- For the first COBWEB implementation prototypes, the use of a remote processing service is considered most appropriate as it simplifies testing and maintenance of conflation functionality. However, if there is no connectivity in the field, the conflation process needs to be performed subsequent to the upload of observations to the COBWEB portal.

- The final application may combine the use of a simple processing library on the mobile device and a remote processing service to avoid internet connectivity issues. Whereas the library is used for basic validity and quality checks, a remote service is applied for advanced quality control and complex processing tasks subsequent to the upload of observations to the COBWEB system. Furthermore, remote processing enables the use of restricted data for conflation, because it is not sent to the mobile device.

- Current WPS frameworks require a lot of processing capabilities and have several dependencies to other software packages. Accordingly, an implementation on a mobile device demands a lot of efforts. Thus, current research on transferring a WPS to a mobile device indicates this may not be the path to follow.

### 3.5. Available data sources for conflation

For the identification of eligible data sources for the conflation with crowdsourced environmental observations, a short survey was carried out among the case study leaders. The results are summarised in Table 2. However, if appropriate, additional data sources may be used during the COBWEB project.

#### Table 2: Possible data sources for conflation

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INSPIRE</strong></td>
<td></td>
</tr>
<tr>
<td>Format: INSPIRE data download services</td>
<td></td>
</tr>
<tr>
<td>Content: as specified by INSPIRE Annexes, for example administrative units, reference grid, protected sites, environmental monitoring facilities</td>
<td></td>
</tr>
<tr>
<td><strong>GEOSS</strong></td>
<td></td>
</tr>
<tr>
<td>Access point: GEOSS Portal (<a href="http://www.geoportal.org">http://www.geoportal.org</a>)</td>
<td></td>
</tr>
<tr>
<td>Format: link to data services and downloads, primarily raster data</td>
<td></td>
</tr>
<tr>
<td>Content: data on various topics (for example water, ecosystems, biodiversity)</td>
<td></td>
</tr>
<tr>
<td><strong>Global Biodiversity Information Facility, GBIF</strong></td>
<td></td>
</tr>
<tr>
<td>Access point: GBIS Portal (<a href="http://www.gbif.org">http://www.gbif.org</a>)</td>
<td></td>
</tr>
<tr>
<td>Format: REST Service interface, JSON encoding</td>
<td></td>
</tr>
<tr>
<td>Content: Global Biodiversity data and citizen observations</td>
<td></td>
</tr>
</tbody>
</table>

### OS OpenData (GB)

- **Access point**: https://www.ordnancesurvey.co.uk/opendatadownload
- **Format**: download available is various formats (for example Raster, Shape, GML)
- **Access constraints**: conflated datasets need to acknowledge the Ordnance Survey
- **Content**: topographic data, administrative units, gazetteer, terrain

### iRecord (GB)

- **Access point**: project website: http://www.brc.ac.uk/irecord/
- **Format**: csv (inhomogeneous)
- **Access constraints**: login required, not allowed to be stored or used for academic research
- **Content**: wildlife observations made by volunteers

### iSpot (GB)

- **Access point**: project website: http://www.ispotnature.org/
- **Access constraints**: data cannot be obtained as direct download
- **Content**: wildlife observations made by volunteers (geo-tagged images)

### NBN Gateway (GB)

- **Access point**: http://www.nbn.org.uk/
- **Access constraints**: specific terms and conditions for each dataset, permission required for derivation of third-party products
- **Content**: biodiversity information

### Local Record Centres (GB)

- **Access point**: no central access point
- **Access constraints**: might not be obtainable for free
- **Content**: biological data

### Marinedateninfrastruktur, MDI-SH (GER - SH)

- **Access point**: http://mdi-sh.org
- **Format**: data services (WFS, WMS)
- **Content**: authoritative data on coastal area of SH (for example species distribution, gazetteer, harmful substances)

### Geodata.gov.gr (GR)

- **Access point**: http://geodata.gov.gr
- **Access constraints**: conflated datasets need to acknowledge creator of the data
- **Content**: available authoritative data for Greece (for example administrative units, protected areas)

---

This project has received funding from the European Union’s Seventh Programme for research, technological development and demonstration under grant agreement No 308513
4. **Conflation use cases**

This chapter identifies a number of abstract use cases that need to be implemented in order to enable a conflation sub-system within the COBWEB framework. The primary use cases as identified are depicted in Figure 4. Beside the COBWEB system, the main actors are (1) the volunteer as part of the crowd collecting environmental information in the field, (2) the expert or scientist, who uses the conflated data for further research or decision making, (3) the data provider, who provides reference data to be used within the conflation process and (4) the service provider, who provides web services required to match, interlink and conflate spatial data based on certain spatio-temporal or thematic aspects.

![Figure 4: Overview on conflation use cases](image)

The following use case descriptions give a brief summary as well as a detailed description of the use cases following the ‘fully dressed’ template as described by Alexander and Beus-Dukic (2009, p. 121).

### 4.1. **Provide Reference Data**

To enable web-based conflation of environmental observations from the COBWEB project with specific reference data, this data needs to be provided in an open accessible fashion. The use of spatial data services is considered advantageous at this point, and spatial databases or files can be provided as appropriate. In any case the provided data should follow the specifications defined by the corresponding COBWEB case study and be complemented with metadata in accordance to ISO19115, especially on quality information and access constraints.

---

This project has received funding from the European Union’s Seventh Programme for research, technological development and demonstration under grant agreement No 308513
Summary: Spatial data is published by a data provider and accessed by the COBWEB online system for conflation with crowdsourced observations

Actors:
Data provider: Wants to provide spatial data
COBWEB system: Wants to use the provided data for conflation purposes

Main Scenario:
1. The data provider collects (feature, coverage or observation) data for publishing
2. The data provider makes the data available online via an open standardised interface
3. The data provider registers the data in an online catalogue with sufficient metadata on data quality and access constraints compliant with ISO19115 and, if applicable, INSPIRE specifications.
4. The COBWEB system searches and requests the provided data

Variation Scenarios: none

Exceptions:
2a. The data is not provided via an open standardised interface:
   2a1. A mediation component is implemented for the COBWEB System to access the provided data
2b. The data is not available online
   2b1. The data is directly included in the COBWEB system to allow for offline access on the server-side; if this is not feasible, the COBWEB system cannot make use of the provided data
3a. The data is not registered in a catalogue
   3a1. The data is directly added or connected to the COBWEB online system
3b. The data is not complemented with sufficient metadata
   3b1. Required metadata is generated or assumed by the COBWEB system
4a. The COBWEB system cannot access the provided data due to privacy, licensing or other access restrictions
   4a1. The COBWEB system cannot make use of the provided data

Preconditions:
Data is available, quality assured and prepared for publication;
Trigger: mandated or voluntary provision of data

Postconditions:
Success guarantees: The provided data is online available through an open standardised interface with sufficient metadata according to ISO19115
Minimal guarantees: The provided data is available to the COBWEB system for further use within the conflation sub-system

Constraints: if the provided data is part of the INSPIRE Annexes I-III, the data provider should be the corresponding authoritative organisation

This project has received funding from the European Union’s Seventh Programme for research, technological development and demonstration under grant agreement No 308513
Provide Conflation Services

To enable web-based conflation in COBWEB, web services offering corresponding functionality are required. To facilitate interoperability and flexible composition, all services should use open standardised interfaces and provide sufficient metadata in accordance to ISO19119. The inputs and outputs used by the service as well as its provided functionality should be formalised and encoded using open standards, if available. For the provision of INSPIRE relevant data, the service provider should be the corresponding authoritative organisation to comply with the subsidiarity principle of SDI.

Summary: Services for the conflation of spatial data are published by a service provider and accessed by the COBWEB online system.

Actors:
Service provider: Wants to provide conflation services
COBWEB system: Wants to use the provided services

Main Scenario:
1. The Service provider implements functionality for the conflation of spatial data
2. The Service provider makes the functionality available online via an open standardised interface, such as the OGC WPS
3. The Service Provider registers the functionality in an online catalogue with sufficient service metadata according to ISO19119
4. The COBWEB system searches, finds and binds the provided service

Variation Scenarios: none

Exceptions:
2a. The service functionality is not provided via an open standardised interface:
   2a1. A mediation component is implemented for the COBWEB System to access the functionality
2b. The service functionality is not available online
   2b1. The functionality is directly coupled to the COBWEB system to allow for offline access on the server-side; if this is not feasible, the COBWEB system cannot make use of the provided functionality
3a. The service functionality is not registered in a catalogue
   3a1. The functionality is directly added or connected to the COBWEB online system
3b. The service functionality is not complemented with sufficient metadata
   3b1. Required metadata is generated or assumed by the COBWEB system
4a. The COBWEB system cannot access the service functionality due to privacy, licensing or other access restrictions
   4a1. The COBWEB system cannot make use of the service functionality
Preconditions:
Trigger: mandated or voluntary provision of service functionality for conflation

Postconditions:
Success guarantees: The service functionality for conflation is online available through an open standardised interface with sufficient metadata according to ISO19119
Minimal guarantees: The functionality is available to the COBWEB system for further use within the conflation sub-system

Constraints: The service provider can be the COBWEB system

Collect and Upload Data
The aim of the COBWEB project is to encourage volunteers to provide environmental data from Biosphere Reserves to the COBWEB online system in order to use this data for the enhancement of existing data, further research or decision making. Therefore, citizens need to collect and upload their observations from the field. Under certain circumstances, which need to be specified by each case study, this upload or additional quality assurance processes trigger a conflation process, which combines the uploaded observations with existing datasets or additional observations from other users.

Summary: Citizens collect and upload environmental observations from the field to the COBWEB online system.

Actors:
Citizen: Wants to collect and provide environmental observations
COBWEB system: Wants to collect and store the observations

Main Scenario:
1. The citizen collects data in the field
2. The citizen uploads its observations to the COBWEB online system
3. The COBWEB system validates the observations
4. The COBWEB system stores the observations
5. The citizen is notified of the stored observations by the COBWEB system

Variation Scenarios: none

Exceptions:
3a. The observations fail the validation process
   3a1. The data is not stored by the COBWEB system
   3a2. The citizen is notified of the failure

Preconditions:
The citizen is sufficiently equipped to collect observations in the field
The citizen is able to upload data to the COBWEB online system
Trigger: the citizen would like to provide field observations to the COBWEB system

This project has received funding from the European Union’s Seventh Programme for research, technological development and demonstration under grant agreement No 308513
Conflate Data

To enhance existing datasets or observations from the surrounding with additional crowdsourced observations taken by the citizen, a conflation of data can be performed. This includes data interlinking and an application-specific combination of data to get an enhanced view on the observed object or phenomenon for further decision-making. Therefore, the process should support the handling of input quality information compliant to ISO19115/19157 and provide corresponding quality information for the outputs.

Summary: Citizen observations are conflated with spatial reference data

Actors:
COBWEB system: Wants to perform the conflation process on collected data (periodic)
Expert: Wants to perform the conflation process on collected data (one-time)

Main Scenario:
1. The COBWEB system (or system component) automatically triggers the conflation process
2. The interlinking process is executed
3. The links are obtained from the interlinking process (see Use Case ‘Interlink Data’)
4. The linked features are application-specifically combined to form conflated data
5. The COBWEB system stores the conflated data for further use

Variation Scenarios:
1a. The expert manually triggers the conflation process

Exceptions:
4a. An error occurs during the conflation process
   4a1. The process stops and returns an error log

Preconditions:
Data for conflation is available and accessible to the COBWEB system
Functionality for conflation is accessible to the COBWEB system
Trigger: the conflation of data is requested

Postconditions:
Success guarantees: The data is conflated and stored by the COBWEB system
Minimal guarantees: The result of the conflation process is returned
Interlink Data

Data interlinking is performed during the conflation process, but can also be viewed separately. Whereas conflation targets the extraction of new value-added information from existing data, the interlinking process connects data based on their specific spatio-temporal or thematic characteristics without drawing any conclusions. Thus, it primarily includes feature similarity measurements, feature matching and the evaluation of feature relations. The results should be encoded and stored in a standardised manner to enable their use within the conflation process or for direct access by clients. Therefore, the use of the Linked Data paradigm is recommended.

Summary: Citizen observations are linked with spatial reference data

Actors:
COBWEB system: Wants to perform the interlinking process

Main Scenario:
1. The COBWEB system triggers the interlinking process
2. The required data (observations, reference data etc) is obtained (see Use Case 'Access Reference Data')
3. The data is matched based on certain spatio-temporal or thematic characteristics
4. The COBWEB system stores the identified links, following the Linked Data paradigm

Variation Scenarios: none

Exceptions:
2a. Data cannot be obtained as required
   2a1. The process stops and returns an error log
3a. An error occurs during the matching process
   3a1. The process stops and returns an error log
4a. The interlinked data is not encoded follow the Linked Data paradigm
   4a1. A mediation component is implemented for the COBWEB system to enable the decoding of the interlinks

Preconditions:
Data for interlinking is available and accessible to the COBWEB system
Functionality for interlinking is accessible to the COBWEB system
Trigger: the interlinking of data is requested

Postconditions:
Success guarantees: The data is interlinked and stored by the COBWEB system following the Linked Data paradigm
Minimal guarantees: The result of the interlinking process is returned

Constraints: none
Access Reference Data

To enable web-based conflation of spatial data, the required input data needs to be accessible by the COBWEB system. Afterwards, it can be combined with uploaded citizen observations. However, the access to reference data can be subject to a number of restrictions, due to specific license agreements, terms and conditions, data structures or provision means.

Summary: reference data is accessed for use within spatial data conflation

Actors:
COBWEB system: Wants to access the provided data
Data provider: Wants to provide reference data

Main Scenario:
1. The COBWEB system searches spatial reference data for conflation
2. The COBWEB system identifies suitable data for conflation
3. The COBWEB system requests the previously identified data
4. The data provider returns the requested data to the COBWEB system for further use

Variation Scenarios: none

Exceptions:
2a. The COBWEB system cannot find suitable data for conflation
   2a1. No data is returned
3a. Access restrictions apply to the reference data
   3a1. The user is requested to provide credentials
   3a2. The data is accessed using the provided credentials
   3a2a. If authorisation is not successful, no data is returned.

Preconditions:
Reference data is provided and accessible to the COBWEB system
Trigger: reference data for conflation is requested

Postconditions:
Success guarantees: Reference data is returned and stored by the COBWEB system
Minimal guarantees: The result of the data search is returned

Constraints: none

Access Conflated Data

To facilitate the further use of conflated data for validation, further processing or research, it needs to be provided in its original form. Therefore, the use of open standards is recommended to facilitate interoperability. To comply with data privacy regulations, proper authorisation and authentication means need to be implemented.

This project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No 308513
in accordance with corresponding case study requirements. If part of a quality assurance process, the conflation may help to assess the quality of inputs.

**Summary:** conflated data is accessed for further analysis and use

**Actors:**
- COBWEB system: Wants to provide the conflated data
- Consumer: Wants to use the conflated data

**Main Scenario:**
1. The COBWEB system provides access to previously conflated data via an open standardised interface
2. The consumer searches for existing conflated data
3. The consumer identifies suitable data
4. The consumer requests and accesses the provided data

**Variation Scenarios:**
1a. The COBWEB system provides access to conflated data for direct download

**Exceptions:**
2a. The consumer cannot find suitable data
   2a1. No data is returned
   2a2. A request can be made by the consumer, to request the COBWEB system for specific conflation results
4a. Access restrictions apply to the conflated data
   4a1. The user is requested to provide credentials
   4a2. The data is accessed using the provided credentials
   4a2a. If authorisation is not successful, no data is returned.

**Preconditions:**
Conflated data is provided and accessible within the COBWEB system
Trigger: consumer wants to access conflated data

**Postconditions:**
Success guarantees: The consumer can access the conflated data for further analysis and use
Minimal guarantees: The result of the conflated data search is returned

**Constraints:** none

**Visualise Conflated Data**
If conflated data is restricted and cannot be provided in a spatial data format, a visualisation of data on a map might still be allowed. This use case specialises the “Access Conflated Data” use case and provides visualisation of the conflated data, ideally through an open standardised mapping interface, such as the OGC WMS. If applicable, the visualisation should include information on the quality of conflated data. This may include geometric deviations from the inputs, attribute differences or...
uncertainty visualisations as described by MacEachren et al. (2005). In accordance with case study requirements, certain access restrictions might apply to the visualisation.

**Summary:** conflated data is visually prepared to be presented as a map

**Actors:**
- COBWEB system: Wants to provide a visualisation of the conflated data
- Consumer: Wants to see the conflated data in a map display

**Main Scenario:**
1. The COBWEB system provides access to previously conflated data via an open standardised mapping interface (for example OGC WMS)
2. The consumer searches for existing conflated data maps
3. The consumer identifies suitable maps
4. The consumer requests and accesses the provided maps

**Variation Scenarios:**
1a. The COBWEB system provides access to conflated data maps for direct downloadable

**Exceptions:**
- 2a. The consumer cannot find suitable maps
  - 2a1. No data is returned
  - 2a2. A request can be made by the consumer, to request the COBWEB system for specific conflation maps
- 4a. Access restrictions apply to the conflated data visualisation
  - 4a1. The user is requested to provide credentials
  - 4a2. The data is accessed using the provided credentials
  - 4a2a. If authorisation is not successful, no data is returned.

**Preconditions:**
- Conflated data maps are provided and accessible within the COBWEB system
- Trigger: consumer wants to access conflated data maps

**Postconditions:**
- Success guarantees: The consumer can access the conflated data maps for visualisation
- Minimal guarantees: The result of the conflated data map search is returned

**Constraints:** none
5. Conflation architecture

For the design of the COBWEB architecture for conflation, a number of dependencies have to be taken into consideration. Essentially, these are the COBWEB functional architecture (D3.1), the mobile data collection sub-system (D4.1), the quality assurance sub-system (D4.2), the privacy assurance framework (D5.1) and the data management guidelines (D7.1). However, the interaction between sub-systems will be based on open standards for data encoding and service provision. Thus, conflation components can be seen as independent, but loosely coupled to the COBWEB system.

Service components

The core components of the conflation sub-system within COBWEB are depicted in Figure 5 and comprise the COBWEB system and additional conflation services. Although defined separately, conflation services can be accessed by COBWEB system components, such as quality assurance, middleware or mobile clients, and in turn also considered as a part of the COBWEB system. Although not explicitly mentioned, proper authorisation and authentication needs to be assured. Therefore, if applicable, the services will participate in the access federation as designed for the COBWEB framework. Please refer to D5.1 for further information on the security and privacy aspects.

Figure 5: Conflation core components for COBWEB

Data Access Services are the main access point for the retrieval of crowdsourced observations and corresponding authoritative reference data from the COBWEB system. Those can be used as input for the conflation process and should offer data...
via open standardised interfaces. Therefore, the use of OGC interfaces standards for
data download, in particular the WFS and SOS, are considered as most appropriate.
Ideally, the provided data is already quality assured and complemented with sufficient
metadata on quality, provenance and lineage.

The COBWEB Registry can be used to register and find suitable datasets or data
services as input for conflation. Furthermore, conflation services and results can be
registered as well, to facilitate the access to conflation functionality and data for
COBWEB users. The corresponding OGC standard that should be used for interacting
with the registry is the CSW.

Quality Assurance Services represent a COBWEB sub-system that can be used to
support conflation processes, in particular the data enhancement process. On the
other hand, those services can access conflation functionality to support quality
assurance for crowdsourced observations by conflating them with validated, most
likely authoritative, data sources.

Conflation Process Services provide conflation functionality to the Web and in
particular the COBWEB system. The OGC WPS standard is suggested to be applied
to conform to open standardisation. The services can implement basic, high-level and
application-specific conflation functionality and can also act as a service orchestration
layer. As a result, the conflation processes returns feature relations or an actual
conflated dataset derived from the input sources, encoded in a standardised format
that can be interpreted by COBWEB clients.

Conflation Visualisation Services can be implemented on top of the processing
services to provide visualisations of conflated data. This can be useful, if clients only
support visualisation interfaces or access to conflated data is restricted. Here, the
OGC WMS can be utilised to act as a wrapper for conflation processing services.

A Linked Data Store is used to store relations between input datasets following the
Linked Data paradigm. Therefore, the relations are encoded in RDF and stored for
subsequent access by clients. The interface to this store can be based on a RDF file
database or a triple store supporting SPARQL.

5.2. Workflow

The actual workflow and results of conflation processes within the COBWEB system
mainly depend on the conflation target and use case requirements. The general
scenario is depicted in Figure 6 and comprises:

The collection of environmental observations by the crowd, defining a certain
spatio-temporal extent and type of observation. The data is first uploaded, then
quality assured and finally stored in the central COBWEB repository.

This project has received funding from the European Union’s
Seventh Programme for research, technological development
and demonstration under grant agreement No 308513
The provision of thematically relevant reference data by a competent authority (for example data specified by INSPIRE), matching the spatio-temporal extent of the observations. This data is stored by the responsible data provider and can be accessed by the conflation sub-system.

The conflation process that identifies links and relation types between the input sources. The execution of the process can be either synchronous, triggered by a client request, or asynchronous, to regularly update the relations between the input sources.

The storage of links and relation types between the inputs in a Linked Data repository for access and further use. The underlying data structure of this store depends on the application and can be implemented in different ways. The pure Linked Data approach simply stores relation types and additional metadata. As an extension to this, conflated inputs can additionally be stored, encoded as spatial Linked Data. In addition, a traditional spatial data store can be linked, holding the conflation results.

The request for conflated data (spatial data format or visualisation), which may already include spatial, temporal or thematic filters. Depending on the type of data store, a conflation process is triggered or previously stored conflated data is directly returned to the client, encoded in an appropriate format.

Figure 6: Basic workflow for Linked Data based conflation
However, different workflows may be derived from the previous one to fulfil more specific purposes. Thus, for example, a demanded real-time conflation does not necessarily store conflation results in a Linked Data repository, but might instead directly process input sources and return the result to the client. This approach is well suited for one-time requests on a low amount of data. Another example is a regular update of a Linked Data repository, which can be useful for frequently requested conflation results and large chunks of data. Therefore, relations are updated in regular intervals without any client request.

By applying a Linked Data approach all inputs remain in its original form and, in the case of authoritative reference data, keep its normative status. The resolving of feature relations depends on the application purpose and can be executed on top of the Linked Data store. As an example, a client may request conflated data with a matching confidence of at least 90%. Based on the relation type and confidence information, a result can be compiled on demand from the Linked Data store.

5.2.1. Provenance and Lineage

To facilitate quality, traceability and security aspects of the conflation process, information on provenance and lineage need to be compiled and attached to conflation results. This affects both the original conflated data and its visualisation. Although the need for provenance and lineage information may vary between use cases and applications, it is recommended to implement it as completely as possible.

Provenance information primarily describes the origin of data and the corresponding quality and security information. This is mainly done using the metadata fields specified by ISO19115 or INSPIRE. The aim is to provide sufficient information to enable the access to data sources and to provide sufficient quality information for the inputs and outputs of a conflation. Lineage information, a part of complete provenance, is important for documenting the conflation process and enable the reproducibility of results. This requires a formalised description of processes that were used within a conflation process, comprising access points, used input and output data and the functionality of a process. This is very important for effective error propagation, especially for successive conflation processes, and can considerably enhance the reliability of conflation results. Therefore, certain metadata elements need to be extended or added to transparently and comprehensively describe the outputs of a conflation process or workflow (Henzen et al. 2013).

In a workflow with successive conflation processes, provenance and lineage information can be stored in two ways: (1) the metadata of an output contains all information on previous processes and used datasets up to the original input sources (integrated view) and (2) the metadata of an output contains the information on the current process and links to the input sources, which in turn contain their own lineage and provenance information (traversed view).
Quality control

An effective quality control during the conflation process requires efforts on quality formalisation and capabilities for the processing of quality information. Aligned to D4.2, a possible interaction between conflation and quality control processes is depicted in *Figure 7*. Although a number of alternative workflows are also possible, it addresses the basic principles. The main workflows are:

1. **Conflation**: Subsequent to the request for conflation, suitable data is searched and retrieved from the COBWEB system and, if applicable, external data sources. If the data does not contain sufficient quality information, a quality control process is requested and performed. Finally, the conflation process is performed, using the quality metadata produced during the quality control process.

2. **Quality Control**: Similar to the conflation process, the quality assessment starts with the search and retrieval of data from the COBWEB system or external data sources. If reference data for conflation is available a conflation process is triggered to add conflated information on the input sources. Finally, the quality assurance is carried out.

During both processes, data or metadata is generated and needs to be stored accordingly. This can be done by using a persistent or temporary data store or attaching the generated data to the inputs. A persistent data store needs to be provided by the COBWEB system to enable a permanent storage of intermediate or final results of a conflation or quality assurance process. This information can later be reused, if appropriate. A temporary data store can be provided by either the COBWEB system or the implemented service components. However, the data will be deleted after the process finishes and cannot be reused. If feasible, the data can also be attached to the input sources and stored in the source data repository. Although this might be considered as the ideal solution, it will not be possible in many cases, especially when dealing with authoritative data.
This project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No 308513
6. Conclusion

As taken from the literature research, quality is one of the most important aspects for the conflation of crowdsourced and authoritative data. Since low quality hampers matching and integration processes, appropriate quality control mechanisms need to be in place. This includes corresponding descriptions of the conflation inputs and outputs as well as the propagation of quality elements during conflation processes. Furthermore, lineage and provenance information is required to distinguish crowdsourced and authoritative data in the conflation result.

Regarding the composition of service based conflation processes, a multi-levelled approach is favoured. Fine granular services facilitate interoperability and reuse, but should already contain an independent and logical functional block for spatial data processing. Aggregate services can then be compiled to serve high level or application specific functionality. The target is to find a good balance between the flexibility of services and computational performance of the aggregate service. Furthermore, open standardisation is important to facilitate interoperability and maintenance within the conflation sub-system. Therefore, the use of OGC standards, in particular the Web Processing Service specification, is envisaged. Attention is also drawn on the current development of the WPS 2.0 standard, which is expected to be released during the COBWEB project lifetime.

The Linked Data paradigm is identified as a suitable approach to bridge current SDI developments with the increasing amount of volunteered spatial information on the web. It can be used to keep the input sources separated and enables a customised conflation based on specific user or case study requirements. Albeit spatial Linked Data will not replace established SDI as implemented for INSPIRE, it is expected to significantly grow within the next years, mainly driven by Open Data movements. To which extent Linked Data can be used for the storage of feature relations and conflation results mainly depends on the particular application needs and thus, needs to be elaborated independently for each COBWEB case study.

The conflation sub-system for COBWEB will be based on the architecture and specifications described in this deliverable. In particular, all system components will adhere to open standardisation, functional decomposition as services and Linked Data principles. The first release will be available for D4.4 “Mobile data collection and validation system software v1” (M24) and the final release for D4.6 “Mobile data collection and validation system software v2” (M36). The documentation of the final system will be included in COBWEB D4.7 “System Documentation” (M42).
7. References


This project has received funding from the European Union’s Seventh Programme for research, technological development and demonstration under grant agreement No 308513


This project has received funding from the European Union’s Seventh Programme for research, technological development and demonstration under grant agreement No 308513


This project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No 308513


This project has received funding from the European Union’s Seventh Programme for research, technological development and demonstration under grant agreement No 308513
8. **Appendix 1: Document History**

<table>
<thead>
<tr>
<th>Author</th>
<th>Version</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stefan Wiemann</td>
<td>v0.1</td>
<td>Structure of the document</td>
</tr>
<tr>
<td>Stefan Wiemann</td>
<td>v0.2</td>
<td>Revision of the structure; added content to several chapters</td>
</tr>
<tr>
<td>Stefan Wiemann</td>
<td>v0.3</td>
<td>Revision of the structure; added content to several chapters, incorporation of comments by UNOTT (SM, DL)</td>
</tr>
<tr>
<td>Stefan Wiemann</td>
<td>v0.4</td>
<td>Added content to several chapters. Contributions by LB and PK</td>
</tr>
<tr>
<td>Stefan Wiemann</td>
<td>v0.5</td>
<td>Added chapter on available authoritative data for conflation</td>
</tr>
<tr>
<td>Stefan Wiemann</td>
<td>v0.6</td>
<td>Minor revision. Addressed comments from DL</td>
</tr>
<tr>
<td>WM/CH</td>
<td>v1.0</td>
<td>Minor edits prior to submission</td>
</tr>
</tbody>
</table>

This project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No 308513