

***Deliverable 2.3: Data Flow***

|  |  |
| --- | --- |
| Document Title | Deliverable 2.3: Data Flow |
|  |  |
| Due date | 31/10/2015 |
| Date of submission |  |
|  |  |
| Lead beneficiary | Hellenic Centre for Marine Research |
| Responsible | FORTH/ICS/ISL |
| Dissemination Level\*1 | RE |
| Nature | Deliverable |
| Status of the Document | 1st Draft |
| Version | 1.0 |
|  |  |
|  |  |

|  |  |
| --- | --- |
| **\*1 Dissemination Level** |  |
| **PU** | Public |
| **PP** | Restricted to other programme participants (including Commission services) |
| **RE** | Restricted to a group specified by the consortium (including Commission services) |
| **CO** | Confidential, only for members of the consortium (including Commission services) |

|  |  |  |  |
| --- | --- | --- | --- |
| **Editor(s):** | **Name** | **Organisation** | **Email** |
|  | Chrysoula Bekiari | FORTH | bekiari@ics.forth.gr |
|  | Martin Doerr | FORTH | martin@ics.forth.gr |
|  | Nikos Minadakis | FORTH | minadakn@ics.forth.gr |
|  | Athina Kritsotaki | FORTH | athinak@ics.forth.gr |
|  |  |  |  |
| **Contributor(s):** |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Table of Contents

[Lifewatch Greece Infrastructure Architecture 4](#_Toc417902921)

[Directory Service 5](#_Toc417902922)

[Metadata Repository 5](#_Toc417902923)

[Data Storage 5](#_Toc417902924)

[Data Flow Workflow 6](#_Toc417902925)

[User Roles 9](#_Toc417902926)

[Data objects 10](#_Toc417902927)

[Mapping processes 11](#_Toc417902928)

[Mapping Definition 12](#_Toc417902929)

[Metadata Transfer 13](#_Toc417902930)

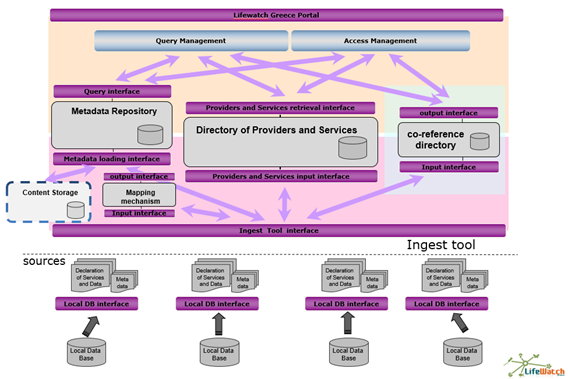
[Change Detection and Update Processing 14](#_Toc417902931)

[Deliverable 2.3 Report 16](#_Toc417902932)

Deliverable 2.3 is partly overlapping with the ***Task 2.2: Data Providers Services Platform***. This deliverable is about the design of the data (and metadata) flow workflow through the Lifewatch Greece infastrastructure. The actual flow will start with the implementation and application of the deliverables 2.2 services.

# Lifewatch Greece Infrastructure Architecture

In order to fulfil the main functional requirements a set of functional components needs to be developed. The data providers provide their metadata (and data) by using the ingest tool, the metadata is mapped to the centralized schema of the infrastructure and is stored in the metadata repository. Information about the way to access the datasets and communicate with the creators/curators is stored into the directory and both data and metadata are stored into the content storage. When all the metadata is stored in the infrastructure the users can execute queries to gain information on how to access the datasets. These queries are handled by the query manager while access manager is responsible for providing access to the metadata repository and the directory. The Architecture is displayed in the figure below and a description of each component follows. More Information about the components of the architecture are provided into deliverable ***2.4 Data Services.***



**Figure 1**. Lifewatch Greece Architecture

## Directory Service

The Directory Service supports the discovery of registered resources within an information community and returns information that allows a user to locate and access the resource and its curator/creator.

Specifically it provides means to:

* Create
* Edit
* Update
* Search

information about providers, collections (datasets) and the means to communicate and access.

## Metadata Repository

The **Metadata Repository** is a virtuoso triple store store where metadata information extracted from the metadata files gets stored as triples. The metadata repository is responsible for:

* Storing
* Updating
* Querying

of the metadata that the providers would like to be public. All the metadata that is provided by the data providers is stored into the metadata repository, after they are mapped to the centralized schema of the infrastructure. The queries that are executed by the users are applied to the metadata repository and datasets that contain the related information are returned.

## Data Storage

The data storage will store any raw data included in the datasets that the Data Providers will upload to the system. This includes MicroCT images, occurrence records, dna sequencing file etc. The data storage is placed in HCMR, but can also be federated to the provider institutions.

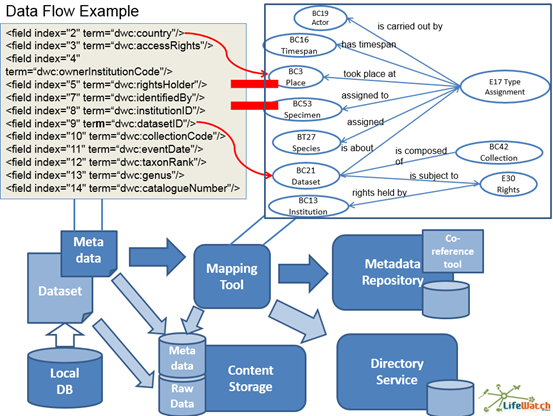
# Data Flow Workflow

The data flow workflow was design based on Synergy Reference Model(SRM),[[1]](#footnote-1) which has been specialized for Lifewatch Greece for the needs of Lifewatch Greece Aggregator (LGA).

The data flow - workflow is briefly described below:

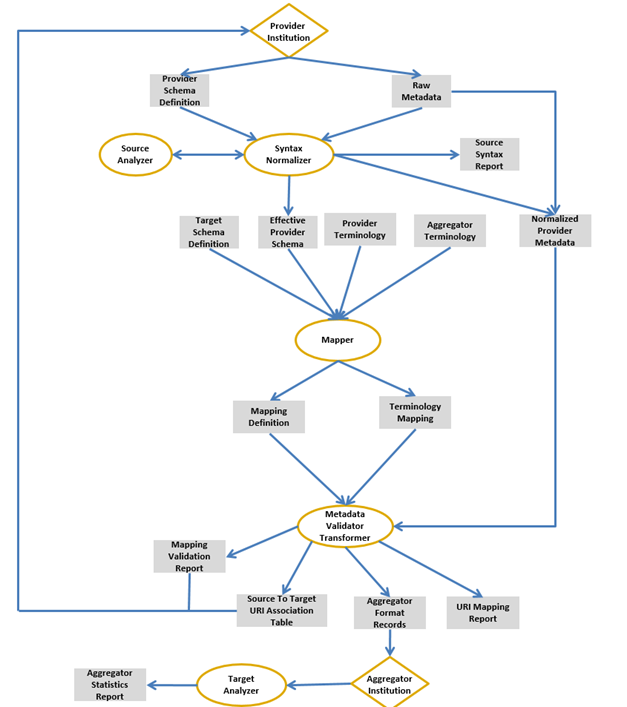
1. Data and Service Providers declare themselves as providers.
2. Providers declare the kind of data they wish to provide.
3. They define and provide the metadata.

1. They provide their content objects (raw data).
2. All metadata is mapped to a unified and normalized XML schema.
3. All metadata is stored in the metadata repository and a part of it in the directory service.
4. Content objects are stored in the content storage
5. The system traces and retrieves terms from the metadata and co-references it with external authorities (Optional Step).
6. End user queries the data/metadata/services by using predefined queries or query formulation.



**Figure 2**. Data Flow High Level Diagram.

A more detailed description of the data flow is depicted in Figure 3. The provider institution gives as input its raw metadata and the provider’s schema definition to the syntax normalizer which produces as output the effective provider schema and the normalized metadata. The effective provider schema is mapped with the Lifewatch’s Infrastructure centralized schema and the appropriate URIs are produced. The mapping definition and the normalized provider metadata are provided as input into the metadata validator transformer which is responsible for the validation and the transformation of the normalized metadata to the Lifewatch Greece’s centralized schema format. The Lifewatch Greece’s centralized schema format records are imported into the Infrastructure.



**Figure 3**. Data Flow Diagram based on SRM

A specification of the user roles and the data object as they came out by the Lifewatch Greece Infrastructure requirements follows:

## User Roles

Primary user roles are:

**Provider:** In Lifewatch Greecewe call providers or provider institutions the providers of biodiversity related data/metadata. We also call them simply *source systems*. Following CIDOC CRM v5.0, “These are either collection management systems or content management systems that constitute institutional memories and are maintained by an institution. They are used for primary data entry, i.e. a relevant part of the information, be it data or metadata, is primary information in digital form that fulfils institutional needs”. In practice and even more general, these are individuals, biodiversity research institutions, museums, universities etc., represented by their curators, IT referents or researchers. Providers *ultimately* have the knowledge about the meaning of their data in the real world (if anybody has it), or know who knows, or know how to verify it.

**Aggregator:** We call aggregator or aggregator institution the maintainer of *the* *target system*. Following CIDOC CRMv5.0, “These provide an homogeneous access layer to multiple local systems. The information they manage resides primarily on local systems.” Aggregator maintain a form of business agreements with providers to send data from local systems to the aggregators’ systems, primarily metadata. In the Case of Lifewatch Greece the aggregation institution is the HCMR where the central metadata catalogue is hosted.

**Source Schema Expert:** The curator, researcher or data manager of the local system who is responsible for the semantically correct data entry into the local system, i.e., the one who knows how fields, tables or elements in the schema correspond to the reality described by them following *local* use and practice. There are a number of source schema experts/data manager placed to the majority of the provider institutions.

**Target Schema Expert:** The expert(s) for the semantics of the schema employed by the aggregator (“integration model”). In the case of Lifewatch Greece this model is the CIDOC CRM and extensions of it such us CRM-sci, CRM-dig and MarineTLO. The target schema experts should include or be in close contact with a sort of curators of the integrated access system to fulfill this role. The target schema experts of Lifewatch Greece are the data managers that are hired and placed in HCMR.

**URI Expert:** The expert of the aggregator, normally an IT specialist, who is responsible for maintaining the referential integrity of the (meta)data in the integrated access system and who knows how to generate from provider data valid URIs for the integrated access system. The URI experts of Lifewatch Greece have defined a number of URI policies that are used by the URI generator functions of Lifewatch’s Data Services.

**Mapping manager:** The actor responsible for the maintenance of the data transformation process from the provider format to the aggregator format. This role may split into a semantic and a technical part, and may be regarded as either an aggregator task, a provider task or a user consortium’s task. The mapping rules are produced by the cooperation of the source and target schema experts, and are used by the Lifewatch Greece Data Services transformation functions.

**Source Terminology Expert:** The curator, maintainer or other expert of one of the terminologies which the provider use as reference in the local system. There is the intention to add this kind of experts in Lifewatch Greece.

**Target Terminology Expert:** The curator, maintainer or other expert of one of the terminologies which the aggregator uses as reference in the integrated access system. There is the intention to add this kind of experts in Lifewatch Greece.

## Data objects

The following primary data objects are distinguished:

**Target system:** i.e., *integrated access systems* in the sense of the CIDOC CRM v5.0 (“These provide an homogeneous access layer to multiple local systems. The information they manage resides primarily on local systems”), to which (meta)data are sent on a regular base or in a single action by several providers. The target system is the Lifewatch Greece Infrastructure which consists of the Directory Service, the Metadata Repository and the Content Storage.

**Content objects:** Individual files or information units with an internal structure that is not described in terms of schema elements of the source or target systems. These vary from MicroCT images and occurrence records to DNA sequences.. They are described as objects by metadata records which are searched by associative queries. Many aggregators do not collect content object but only link them back to the provider system. In Lifewatch Greece the content objects. These content objects can be stored in the Content Storage of the infrastructure without this step to be mandatory. The users have the option to provide metadata about the content objects and the way to link back to the provider institution or to choose to store the exact content object to the centralized content storage.

**Metadata records:** Information units with an internal structure that isdescribed in terms of schema elements of the source or target systems. In our context, these are often data records describing content object (therefore the term “metadata”), but bad analogy brought the term also into use for data describing physical objects and other historical contexts. Therefore we define it here by the way it is treated in the information system, and not as “data about data”. The metadata records are the common target of submission to aggregators and therefore of transformation from the source to the target schema. Fifteen metadata categories have been identified for the Lifewatch Greece needs corresponding to each biodiversity domain concept (occurrence metadata, microCT metadata etc). These records are ingested into the Metadata Repository of the infrastructure.

**Integration Model:** The definition of the schema of the target system, now mostly an RDFS/OWL knowledge representation model (“ontology”). This model is the CIDOC – CRM and some of its extensions such as CRM*sci*, CRM*dig* and Marine TLO.

**Source systems:** I.e., *local information systems* in the sense of the CIDOC CRM v5.0 (“These are either collection management systems or content management systems that constitute institutional memories and are maintained by an institution”) from which (meta)data are sent on a regular base or in a single action to some aggregator.

These are mostly museums of universities which will have their data automatically harvested by the LGA. There is an intension to add this kind of functionality in the future but it not mandatory.

**Terminologies:** Controlled vocabularies of terms that appear as individual data *values* in the source or target systems. We do **never** use the term “vocabulary” for metadata schemata. Terminologies may be flat list of words or be described and organized in more elaborate structures as so-called “thesauri” or “knowledge organization systems”, the most popular format now being SKOS. Here, we do **not** use the term “ontology”, even if the terminology may qualify as such, as long as its use in this context is to provide *data* *values*. There is an intension to add terminologies as data objects of the LGI but this addition is optional.

**Source schema definitions:** data dictionaries, XML schemata, RDFS/OWL files etc. describing the data structures that are managed and can be searched by associative queries in the source or systems. The most popular source schema definition is the Darwin Core.

## Mapping processes

Lifewatch Greece aims at identifying, supporting or managing the processes needed to be executed or maintained when the provider and aggregator agree:

(1) to ***transfer data*** from the provider to the aggregator,

(2) to ***transform*** their format to the (homogeneous) format of the aggregator,

(3) to ***curate*** the semantic consistency of source and target data and the global referential integrity and

At a first level, this breaks down into the following independent processes:

(a) Management of which data will be delivered and processed at what time, including updates.

(b) The mapping definition, i.e., specification of the parameters for the data transformation process, such that complete sets of data records can automatically be transformed.

(c) The actual transfer of data until a first consistent state is achieved. This includes transformation of sets of data records submitted to the aggregator and ingestion of the transformed records into target system.

(d) Referential integrity processing at the aggregator side out of the context of a particular data submission.

(e) Change detection and update processing to restore ability of data transformation and semantic consistency.

## Mapping Definition

Mapping definition breaks down into

* Syntax normalization
* Schema matching
* URI generation specification
* Terminology mapping

The *Mapping Manager* is responsible for issuing and coordinating these tasks.

#### Syntax Normalization

Some provider systems do not employ the data or database formats of RDBMS, MSExcel spreadsheets, XML or RDF/OWL which are now standard, but may use tables in text documents or other formats. The latter do generally not impose formal control on the syntax corresponding to the intended semantics. Even standard formats allow for the use of free text fields where users may insert their own syntactic inventions without a formal control, such as punctuation marks, italics, parentheses and others. XML further allows for semi-structured data and the use of elements not appearing in a schema declaration. Local identifiers may have their own syntactic structure, such as inventory numbers, addresses, bibliographic references, date and time etc. Automated data transformation, i.e. transformation of data from one schema to another without loss of meaning or with controlled loss of meaning by a deterministic algorithm based on a mapping definition is only possible if the part of the data to be transformed (not the free text descriptions) is completely structured. Therefore, whatever data is to be mapped and transformed has to be completely structured and each structural element must have a clearly described meaning. The syntax normalization in Lifewatch is achieved by a combination of metadata template files and the transformations that are applied by the Lifewatch Data Services.

#### Schema Matching

*Source schema experts* and a *target schema expert* (e.g., CIDOC CRM) define a schema matching, which is documented in a ***schema matching definition*** file. This file is human and machine readable and is the ultimate communication means on the semantic correctness of the mapping. The collaboration between these *experts* must be well organized and is the bottle-neck of the mapping process. In order to do so, all source schema elements must be well understood and mapped to target schema paths. The schema matching for the case of Lifewatch Greece is achieved by exploiting the x3ml language capabilities and the functionalities of Lifewatch Greece Data Services.

#### URI Generation Specification

After the *schema matching* process, the URI generation policies for each instance of a *target schema* class referred to in the matching must be defined, such as for persons, objects, events,places, and formats of time. This is typically a task the *source schema expert* has not interest in orknowledge for. The execution of URI generation rules may reveal “dirty data” of the provider at transformation time, or before. “Dirty data filters” must be foreseen in the generation rules. The source metadatarecords may be analyzed before transformation time for such cases. URI generation is achieved by a number of URI generation rules that are executed by the Lifewatch Data Services at the transformation phase of the records.

#### Terminology Mapping

Terminology mapping can be a huge task and there is the intention to be added to LGI data-flow workflow. Providers may use anything from intuitive lists of uncontrolled terms up to highly structured third party thesauri. However, most of the provider terminology is very specialized and more important as information element when metadata records are displayed than as search term in the target system. The aggregator terminology should have a thesaurus structure, albeit a small vocabulary of high-level terms. Sometimes it may be more effective to merge provider terms with aggregator terms, i.e., replace equivalent terms and insert all other provider terms as narrower terms of aggregator terms. In this case, the provider terminology should be *replaced* by the updated aggregator terminology before transformingthe respective records and in the schema matching definition file. This will allow for better controlling the mutual consistency of mappings between different providers. The terminology mapping may reveal “dirty data” of the provider. “Dirty data filters” must be foreseen for terminology. The source metadata records should be analyzed before transformation time for such cases. Providers must be informed about “dirty data” cases, and given the possibility to run an organized, sustainable process to improve the source data.

## Metadata Transfer

#### Submission

In order to transform source metadata, they must be submitted to the aggregator (HCMR Metadata Repository). The submitted metadata records must are intentified with a unique identifier and are kept also to the content storage.

#### Transformation

The transformation process itself runs *completely automatically using the Lifewatch Greece Data Services*. If it encounters dirty data, trusted expertise, *experts* under the control of the *mapping manager* may correct some of them manually in an *interactive process*. The result is a set of valid target records. The URI generation algorithm of the automatic data transformation process may employs an initial instance matching process with the target system in order to reuse already existing URIs in the target system. This also holds for third party authority systems with URIs the aggregator uses as reference.

#### Ingest and Storage

The transformed records are ingested into the target system (Metadata Repository and Directory Service). An aggregator stores all source metadata records which are going to be transformed, or which are transformed and have been ingested to the target system. The aggregator preserves the link to the identity and version of the source record it is derived from. Source records may be syntactically normalized for that purpose. If a new version of a transformed source record is ingested in the target system, the target record representing the previous version must be deleted in the target system if it has the form of a semantic network (without a “context” or “named graph” or “quad store” mechanism, this causes a “data warehouse update problem” at target system in the form of a Triple Store). The aggregator may decide to delete copies of updated source records. Providers should better keep the old versions. If providers have no means to do so (e.g., in RDBMS), aggregators may take over the preservation of old versions as a service.

#### Co-reference Resolution

The aggregator may continue at any time with referential integrity processing, i.e., resolving that multiple identifiers denote the same real world thing or object of discourse (co-reference resolution). There is the intension to add co-reference resolution functionalities in LWI. The main goals are to ensure referential integrity, which is the heart of information integration, and to reduce the number of URIs in use for the same thing. It requires its own dialogue between provider, aggregator and third-party authority managers. Since the aggregator collects more comprehensive knowledge than the providers, it is a natural role of the aggregator. One may regard that the only genuine knowledge of the aggregator is the co-reference knowledge.

## Change Detection and Update Processing

The mapping manager must monitor all changes that may affect the consistency of provider and aggregator data. There is the intention to add such kind of functionality to the LGI.

Those changes are:

1. new source records

2. source record updates (new versions)

3. source schema changes

4. provider changes identifier policy (for people, objects, events, places, time) and updates his records

5. provider changes terminology data (terms or authority) and updates his records

6. provider changes terminology structure (broader term links etc.)

7. target schema changes

8. aggregator changes URI policy

9. aggregator changes terminology (terms or authority)

10. aggregator or user consortium changes mapping guidelines

11. Source-target terminology mapping changes

The changes of kind 11 requires retransformation and re-ingestion of all (stored) source records already transferred which refer to the respective terms. The changes of kind 6 and 9 require redoing the terminology mapping, retransformation and re-ingestion of all (stored) source records already transferred which refer to the respective terms. Changes of kind 8 require updating the URI generation specification in the mapping file, retransformation and re-ingestion of all (stored) source records already transferred which refer to the respective terms. Changes of kind 7 and 10 require updating the schema matching definition, retransformation and re-ingestion of all (stored) source records already transferred which refer to the respective terms. The changes of kind nr. 1 and 2 require running the complete metadata transfer with the changed or new source records but using the existing mapping definition. Changes of kind 3 require updating the schema matching definition in the mapping file, resubmission of all source records affected, transformation and ingestion, replacing the target records transformed from the previous version of these source records. Changes of kind 4 require updating the URI generation specification in the mapping file, resubmission of all source records affected, transformation and ingestion, replacing the target records transformed from the previous version of these source records. Changes of kind 5 require updating the terminology mapping, resubmission of all source records affected, transformation and ingestion, replacing the target records transformed from the previous version of these source records.

# Deliverable 2.3 Report

*Starting Date:* 1/1/2013

*Ending Date:* 31/10/2015

*Deliverable First Version Date:* 31/12/2014

*FORTH Participants:* Athina Kritsotaki: 1/1/2013 – now

Nikos Minadakis: 1/4/2013 – now

Yiannis Marketakis : 1/12015 - now

*Final Version Progress:* 80%

**1/1/2013 – 31/12/2014 Progress Report:**

During this period the functional requirements of the infrastructure were collected, the architecture of the Lifewatch Greece Infrastructure was designed, and every component was defined according to these requirements. The workflow of the data flow was defined and designed, along with the role of each component to it. The Synergy Model was studied, and applied in order to refine the workflow related to the recommended best practices. The first data flow to the infrastructure started.

**Next Steps until 30/9/2015:**

Until the end version of the deliverable the real data flow will begin along with the implementation and application of deliverable 2.2 data services. Any required change to the workflow will take place. A significant number of datasets and metadata is expected to flow to the infrastructure.

1. designed and developed in KRHPIS-POLITEIA (MIS 448300) [↑](#footnote-ref-1)