European Long-Term Ecosystem and Socio-Ecological Research Infrastructure

D8.1 eLTER Information Architecture Report

Authors: John Watkins (CEH), David Ciar (CEH), Christoph Wohner (EAA), Johannes Peterseil (EAA), Herbert Schentz (EAA), Alessandro Oggioni (CNR-IREA), Simone Lanucara (CNR-IREA), Vladan Minić (BSI), Srđan Škrbić (BSI), Žarko Bodroški (BSI), Ralf Kunkel (FZ-Juelich), Jurgen Sorg (FZ-Juelich)

Lead partner for deliverable: NERC/CEH

Other partners involved: EAA, CNR, BSI, FZJ

H2020-funded project, GA: 654359, INFRAIA call 2014-2015

Start date of project: 01 June 2015 Duration: 48 months

Dissemination level

PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	
CI	Classified, as referred to in Commission Decision 2001/844/EC	

Version control	Edited by	Date of revision
Created – V1	Watkins, Bertrand	10.07.2015
Internal review	Schentz, Herbert	19.8.2015
Internal review	Peterseil, Johannes	21.8.2015
Revised – V2	Bertrand, Nicolas	13.6.2016
Revised – V3	Watkins/Ciar	28.7.2016
Internal review	Oggioni, Alessandro	10.8.2016
Internal review	Minić, Vladan	15.08.2016
Internal review	Minić, Vladan	16.08.2016
Internal review	Škrbić, Srđan	16.08.2016
Internal review	Žarko, Bodroški	14.12.2016.
Revised – V4	John Watkins	19.12.2016

Publishable Executive Summary

The eLTER Information System Architecture report lays out the basic components of the IT infrastructure that will be built to support data management, integration and dissemination across the eLTER network. The design of the architecture is based on existing software components and uses international standards to ensure compatibility not only within the network but also with external partners. The architecture will be evolved with users through a number of ‘user stories’. These will be addressed with system users so that the development can be tailored to the eLTER community’s requirements as it is built. This agile approach to the architecture implementation will continue through the project lifetime so that information system will continue to evolve with its increasing use.

Contents

1 Introduction to the eLTER Information System Architecture

1.1 Aim of the document

This document describes the overall information system architecture for the eLTER information system and basic user stories of how its functions will provide benefit to the eLTER community. The eLTER information system will provide the facilities to store, describe, discover, explore, assess and access information on the long term ecological observations and experimentation from across the LTER-Europe site network. The federated nature of the eLTER information system reflects the nature of the community it seeks to support and enables flexibility in creating new partner virtual data nodes (eLTER data nodes), linking to existing partner data nodes using compatible interfaces and openness for others to link eLTER information resources into their own web portals.

1.2 eLTER approach to systems architecture development

The eLTER project is taking an iterative approach to the development of the information system architecture. The different components and links in the system will be developed over time rather than designing a complete system before implementation. The architecture document is an overview of the system architecture as currently envisaged by the project but we expect and hope that this will change as assumptions about the design are challenged by users as they use initial implementations of system components. The process will be driver by a number of “user stories” or test workflows of how we think users will work with the system. These are documented at the end of the report.

1.3 Systems Architecture Overview

The following diagram represents the major components of the eLTER information system architecture and the main connections between these components. This is not a definitive diagram of all components and relationships required to implement the information system as this must necessarily evolve over time as the project iteratively develops the final system with users.

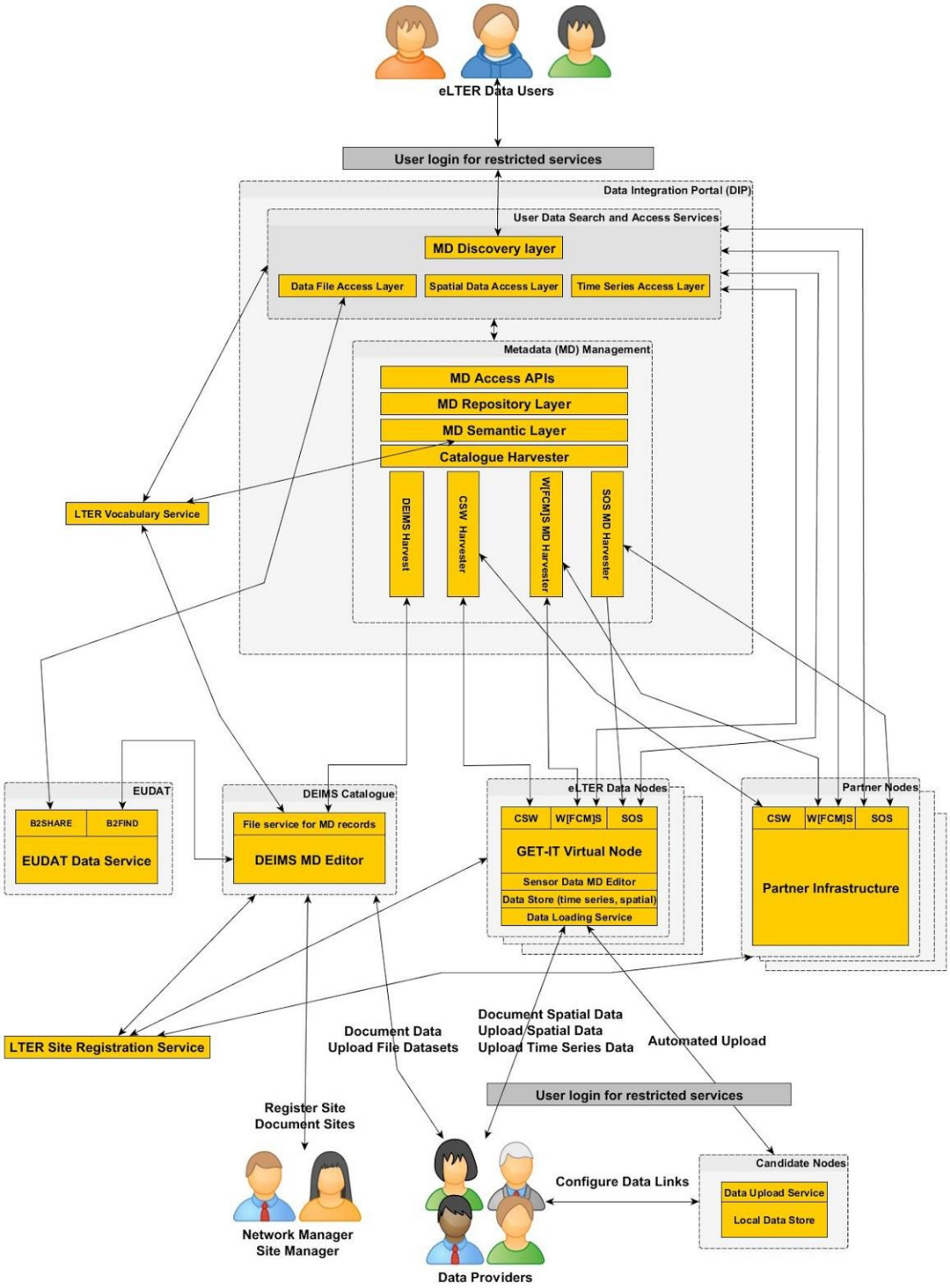

Figure 1: eLTER Information Architecture Overview

2 Components of the system architecture and interface standards

2.1 Data Integration Portal (DIP)

User login and authorisation

All users of the information system will have access to basic browse and access services for unrestricted information. Other services such as administration and access to unpublished data will be through a user login service that will identify users and assign them to roles through a dedicated authentication and authorisation (e.g. LDAP) service. This will be separate from the user accounts of local data portals or the DEIMS metadata editor.

Metadata search module

Metadata search functions will be based on a set of standards implemented through compliant software modules. This will be based on GeoNetwork, 52°North and PyWCS applications in the first instance though other implementations can be accommodated through shared standards (e.g. SOS). These modules will collate information from the different eLTER data nodes in the network and enable users to browse the metadata records and spatial information that describe the data available from the eLTER site network. The harvesting of metadata records from the eLTER data nodes will be based on OGC CSW services providing ISO 19115 standard data set and service descriptions. These records can be browsed in the **MD Discovery Layer** to select either data sets or data services that can then be accessed. Search can be done using map services and time series graphing options through the **Spatial Data Access Layer** and **the Time Series Access Layer**. The 52°North software will be configured to access download from multiple time series data source from the different eLTER sites. Where metadata records point to individual files, these can be accessed via **Data File Access Layer** from the DEIMS file repository and linked EUDAT data services.

Visualization of Sensor data via Sensor Observation Service (SOS) module

Sensor data will be accessed using the OGC SOS standard services between clients in the DIP and services in the eLTER data nodes. A modified set of 52°North SOS modules will be used in the DIP and the eLTER data nodes. Partner data nodes may use other implementation of SOS. This allows access to the multiple eLTER site data sources from eLTER data nodes and partners.

Using 52°North for the DIP Sensor Web modules (e.g. 52°North Helgoland web tier and 52°North REST API) allows additional functions while still supporting partner data nodes. The REST API can be modified so that it supports visualization of data from multiple SOS sources and various options for data filtering. The server supports only caching and visualization of SOS sources that have been defined statically in an XML file, so further modification can be made to allow SOS sources to be added while the server is online. These option will be tested with users and partners during iterative releases.

A visualization module will be integrated with the EnvThes vocabulary server in order to achieve unification of phenomena names that are currently not standardised across the network. The system will support LDAP based authentication of users and options to define public and restricted SOS sources.

Data File Access through EUDAT Integration module

All files uploaded to B2Share repository through DEIMS will be available for end users through the DIP search. Metadata and data search can be done using B2Share or B2Find EUDAT services. Both services use APIs for communication but requests are different e.g unlike B2Find, in order to make B2Share requests, authentication token must be provided. B2 services responds with messages formatted using JSON (JavaScript Object Notation). Those responses will be parsed and shown to users so they can interact with the B2 services from the DIP. Because both services give the same results, in terms of the number of responses and their content, DIP team will decide which service is more suitable for usage and integration.

2.2 Metadata Management System

MD Access APIs

The metadata management system will provide a range of web service interfaces to allow access information from different metadata record types. This allows general queries on particular topics across all eLTER information, e.g. which datasets, sensors or visualisation services are available for meteorological data. These general searches will use standard OGC CSW APIs for communication with the DIP user interfaces and external portals requiring information.

MD Repository Layer

Held in this module are the aggregated metadata records from all the catalogues and data nodes. The metadata records are used by this layer to service the queries of the DIP parsed via the MD API module. These records are harvested by the Catalogue Harvester modules from the link nodes and catalogues. They are collated to a single store of records initially based on the ISO 19115 schema. The semantic consistency in the descriptive terms, properties, and concepts used in the metadata records is mediated through the MD Semantic Layer. This has access to a vocabulary management services, for example the EnvThes service developed by LTER Europe community.

MD Semantic Layer

Semantic mapping is checked within this module, mapping the terms from the harvested metadata, to the terms held in the standard vocabularies which are used by the DIP for discovery. Data is passed from the harvesting services into the MD repository and then this module checks that the terms, properties, concepts, and other metadata values are valid (e.g. from the EnvThes vocabulary). Both eLTER network level descriptors and site specific terms can be supported within EnvThes and linked into metadata descriptions. This provides a way of checking and providing standardised vocabulary term for searches, while at the same time providing the end user, or DIP modules, with the term used in the data node so that the user of the platform know what to search for when they go directly to the data node.

Catalogue Harvester

This module is designed to manage the service calls and harvesting of metadata from data nodes that provide service level access. It has multiple sub-modules, where each of these sub-modules is designed to connect to a specific service type, request all available metadata and place in the MD repository. The different modules are configured to harvest from each registered data node at the required time period. These modules will largely use OGC services such as CSW, WMS and SOS but will also use web file services (e.g. for DEIMS) where necessary.

2.3 Virtual Data Node (aka GET-IT Virtual Node)

eLTER Data Node

The federated nature of the LTER-Europe site network requires the development of a distributed network for sharing of ecological observations and a wide range of ecological data.

The different nodes of eLTER data infrastructure are implemented exploiting the free and open source software suite Geoinformation Enabling Toolkit StarterKit® (GET-IT), developed under the flagship project RITMARE. GET-IT facilitates the creation of nodes of a federated Spatial Data Infrastructure (SDI) for an observational network. The software suite allows users to straightforwardly share their observations and sensor metadata on the web (using OGC standards). It consists of a virtual machine, based on the Ubuntu operating system; the basic software used is GeoNode, a widely known geographic content management system. Sensor metadata creation has been performed by GET-IT metadata editor, called EDI, which allows ease and friendly instrument registration (SensorML editing version v1.0 and v2.0) through graphical user interfaces (GUI) and auto completion facilities linked to vocabularies. In particular, parameters definition have been borrowed from the terms present in the EnvThes environmental vocabulary, in order to harmonize and semantically enrich the metadata with respect to the LTER community.

Sensor Web Enablement (SWE)

The OGC's Sensor Web Enablement (SWE) standards enable developers to make all types of sensors, transducers and sensor data repositories discoverable, accessible and useable via the Web. O&M and SensorML are OGC standard included in SWE framework.

In each GET-IT node the Sensor Observation Service (SOS) functionality is provided by the 52°North SOS application. The SOS provides support for storing, searching, and serving sensor/procedure definition documents and the observations generated by these defined sensor/procedure instances. Within GET-IT the sensor/procedure definitions can be created by EDI metadata editor and stored using version 1.1 or version 2.0 of the SensorML standard. The observations are stored using the O&M 2.0 standard by GUI also present in the GET-IT. The functionality of the SOS can be broken down into three main components of: data entry, search, and retrieval. Each of these components is described in the

paragraphs below, however a common attribute to each component is that generally, the encoding format of the request and result can be one of a selection including: JSON, SOAP, KVP, and POX.

Service Actions

Data entry is provided by the storing and updating of sensor definitions (metadata of the sensor), result templates, and result or observation entries. Sensor definitions are versatile in that they can be defined, and then updated as many times as necessary, with the potential to define certain sensor definitions valid for defined temporal ranges. Sensor definitions also support hierarchy allowing for sensor networks to be defined as linked components, and logical hierarchy in terms of base sensor type and inheritance of this by instance classes of the base type. Lastly, observations can be entered into the SOS, linked to a feature of interest), a procedure, and the observed property. This can be done in a number of ways, but the two standard choices are to either insert each observation individually, or, to create a result template which then allows batches of observations to be added at once using the template to specify the correct metadata with the exception of the phenomenon time, result time, and observed value.

Data search is provided in a very basic sense for sensor definitions, where it is possible to browse all the sensors and procedures defined within the SOS as a list, and to then request a full definition of any particular sensor from that list as desired. Observation or feature data (entered as metadata along with an observation or result template) provides the ability to be searched for by defining optional parameters such as the procedures or sensors that observed them, the type of property being observed, spatial filters, and in the case of observations temporal filters also. An unofficial service call, GetDataAvailability, is highly recommended for use as it provides a summary of data matching the search constraints in a more informative manner than the current OGC compliant requests.

Data retrieval regarding sensors and procedures, as touched upon in the previous paragraph, is simply being provided with the SensorML document that is used to describe them, available by specifying the particular sensor or procedure of interest. Within GET-IT node each sensors metadata can be displayed in two ways: SensorML XML and HTML user friendly presentation layer of SensorML. Observations can be returned as individually encoded entries, with multiple entries in a document, or, as a merged document of observations. It is possible to retrieve a collection of observations specified only by the observation type, or the temporal period, feature of interest, or other search facet, not only by requesting a single sensor or procedure's output.

Observation Representation

Within the GET-IT node, the observations are recorded nominally as single independent observations, as opposed to representation as a discrete time coverage. This distinction is made as the TimeSeries part of the WaterML 2.0 standard provides for the representation of a whole time series, with its own properties and metadata such that any single observation is defined as being within a given time series. With the current SOS module representation, while there can be many observations by the same sensor, in the same location, over a given temporal scale, each observation, by default, has no knowledge or information on the observations that occurred before or after it, and it is the responsibility of the software to

collate and represent the observations as a single time series object.

Dataset METADATA client

The METADATA client is a powerful tool that allows a straightforward and intuitive method to provide the metadata required to create well-formed SensorML and spatial metadata documents. SensorML documents in particular can grow to be both long and complex, increasing the time taken and difficulty in creating, editing, and managing these records. The METADATA client abstracts this away, while at the same time requiring a subset of core data that must be recorded to be INSPIRE compliant and meaningful to other users who may not know or have been previously involved with the data or sensor network being described.

GET-IT CSW Service

An OGC standard catalogue service to handle metadata information requests providing INSPIRE and ISO 19115 discovery of dataset description records. Within GET-IT data node, the catalogue service is implemented using PyCSW.

GET-IT WMS/WFS/WCS mapping services

The OGC standards baseline comprises more than 30 standards including: Web Map Service (WMS), Web Feature Service (WFS), and Web Coverage Service (WCS). These standards web services have been implemented to share different type of geo-data. Within GET-IT data node this services are implemented using GeoServer and drag and drop GUI allows the users to upload raster or shape files. A metadata of this data can be created through the use of EDI metadata editor where terms have been used from the terms present in the EnvThes environmental vocabulary.

GET-IT SOS Data service

Through GET-IT data node users can share their observations. Usually these observations are collected during the field work by data logger or are automatically stored in machine memory; however, generally these are translated into a spreadsheet. Within GET-IT the users have access to different forms that allow them to upload data for specific sensors. The first step allows selection of the sensor that collected the observations, the second allows the user to select the geographic position where the observations are collected, and finally a table to paste in the observations that closes the workflow.

More informations about the GET-IT facilities can be read in the software documentation.

2.4 MetaData Provision module (DEIMS)

The Dynamic Ecological Information Management System (DEIMS) is the current metadata

catalogue used by LTER-Europe to describe the information assets of the network including datasets and field site facilities. DEIMS has editing facilities to enter information for datasets which is then held largely within an Environmental Markup (EML) schema within a Drupal application. The information in this schema can then be transformed into ISO 19115 records for dataset descriptions and INSPIRE EMF records for site descriptions. This allows DEIMS to be used as an authoring tool for dataset and facilities metadata that can then be shared through web service APIs provided by the eLTER Metadata Management System such as CSW. In the first implementation ISO 19115 records will be harvested as XML files from a Web Accessible Folder rather than through a CSW service as this allows greater flexibility given the range of information collected by DEIMS (e.g. site configurations and contacts).

In addition DEIMS provides a forms to document long term observation and experimentation facilities (e.g. LTER Sites or LTSER Platforms) and networks (e.g. national LTER networks). For each site entry a unique identifier, the SITE_UUID is generated (e.g. https://data.lter-europe.net/deims/site/8eda49e9-1f4e-4f3e-b58e-e0bb25dc32a6).

2.5 Links to Champion Partner Nodes

The eLTER information system architecture is designed to be open so that other systems using the same interface standards can supply and consume information. In the eLTER H2020 project, “champion” nodes will provide data through existing OGC web services that they provide to the community (e.g. CSW for metadata and SOS time series data). In order to link these into the eLTER information system, metadata records will be created describing the data, facilities and services offered and linked to the DIP Metadata Management System repository. The metadata records can then be accessed through the search facilities of the DIP to assess the data services available. Using the data services description, the data visualization and download layers can access the Partner Node services to enable integrated data viewing and access alongside other eLTER Data Nodes.

2.6 eLTER Site Registration Service

As part of the architectural services development, a site registration service will be implemented to allow different nodes to create new persistent eLTER site registration identifier codes and then access them in different services such as DEIMS and the DIP. The registration service will be similar to the architecture used within the ICOS ESFRI.

2.7 Vocabulary Services

The eLTER Information System will have descriptions of and access to very varied data that will use a broad range of specific scientific names and vocabularies. Naming conventions evolve over time and so the ability to map names and terminology for similar measurements

and methods is required both across different field sites and also across the historic records of any one site. This mapping of names and reconciling of different terminologies will be carried out through reference to a web based vocabulary service. The eLTER Information System will largely use SKOS-based vocabulary services giving users access to hierarchical naming systems via SPARQL queries. The principal service will be the CEH-hosted vocabulary service at vocabs.ceh.ac.uk which houses the ENVTHES vocabulary developed by LTER-Europe. This service is publically available so can be accessed by any data provides or external users to resolve the meaning and relationships between terms used in the different datasets.

Vocabularies such as ENVTHES will be used via the vocabulary service to describe datasets by various metadata authoring tools such as DEIMS and GET-IT EDI. These terms can then be used to search for datasets that contain similar or related measurements that can be used with confidence in subsequent data synthesis and analysis. This element of semantic mediation between datasets in a key element for bring together data from across the eLTER network and beyond.

3 The initial user stories for implementing the eLTER Information System architecture

As part of an agile development of the eLTER information system, a small number of user stories have been defined to start the development of functionality that users within the project can use. These cover the broad range of functions from a data user and a date provider perspective. The development and testing of components for these use cases will test and modify the architecture specification to ensure it matches user needs

In software development and product management, a **user story** is a description consisting of one or more sentences in the everyday or business language of the end user or user of a system that captures what a user does or needs to do as part of his or her job function. User stories are used with agile software development methodologies as the basis for defining the functions a business system must provide, and to facilitate requirements management. It captures the 'who', 'what' and 'why' of a requirement in a simple, concise way, often limited in detail by what can be hand-written on a small paper note card.

These user stories will evolve in complexity and number over the life of the project to cover all areas of the architecture needed to establish the eLTER information system network while making sure that each cycle of development is relevant to the eLTER user community.

User story 1: As a [data provider], I want [to publish] a [citable eLTER dataset] so that I can securely share my data, meet journal / funding requirements and gain recognition for data outputs

Description: A [data provider] wants to publish and share a defined dataset with the user community. This dataset needs to be citable in order to create credit for the data provider. He/she goes to the LTER Europe node in order to document and reference the information.

The dataset needs to have:

1. a complete MD description using a defined standard used in the community to provide information on the content and use of the dataset
2. a persistent identifier (PID) that can be dereferenced to a metadata data record landing page that allows for download of the specific dataset.
3. a data access services so that other can access the data

Candidate Tasks

1. Enable metadata authoring through the editing tools in DEIMS and assign PID
2. Enable upload of the data set to the EUDAT B2SHARE service
3. Harvest metadata records into the DIP repository
4. Enable search and view of the metadata record in the DIP data discovery layer
5. Enable download of the required dataset via dereferencing PID to the data service link

Candidate Acceptance Criteria

1. Add a metadata record of the required data
2. Upload the metadata record to the PID so that it is accessible from the discovery tools
3. Use the PID as a dereferencable link to the metadata page
4. Download the dataset through the DIP services

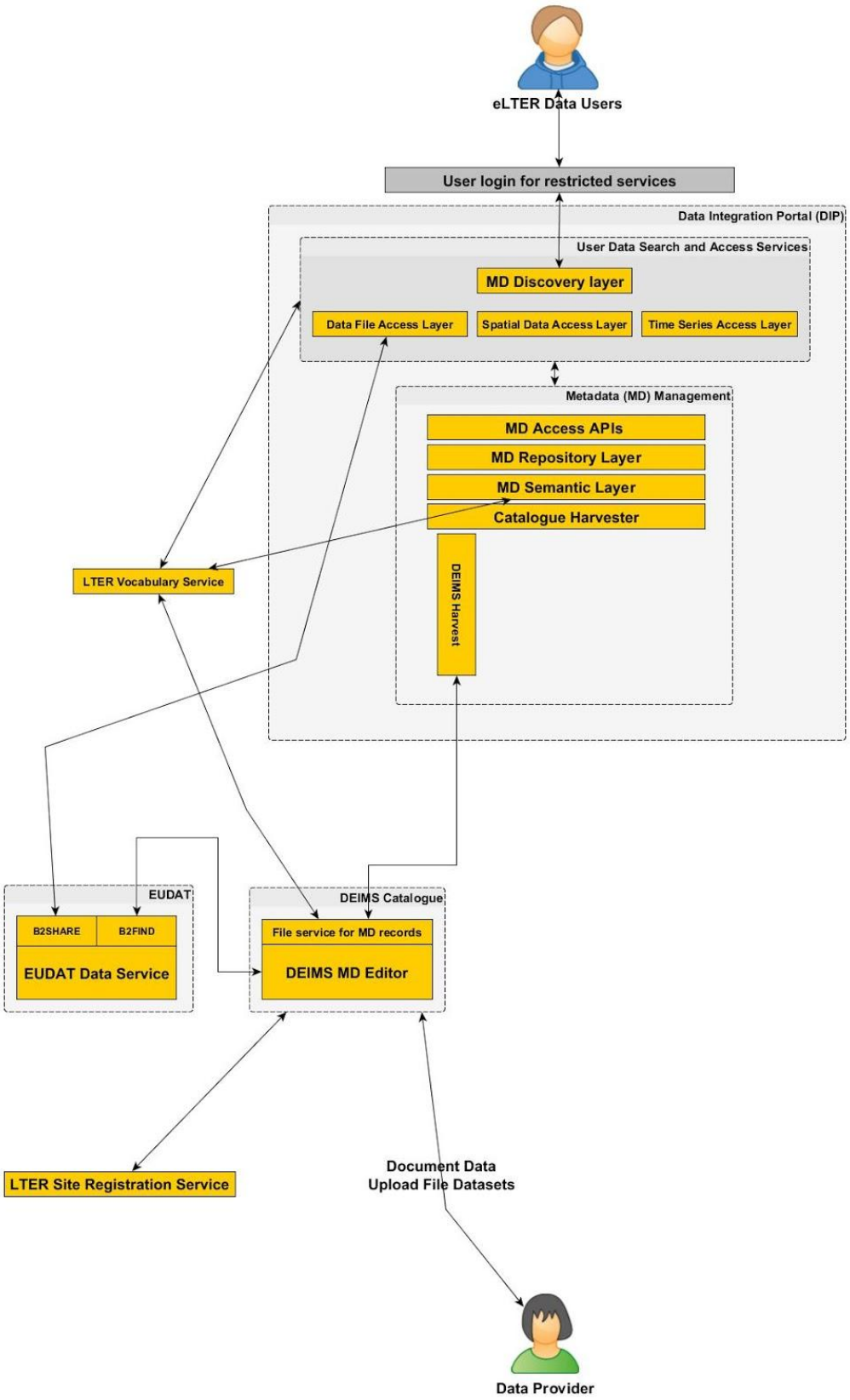

Figure 2: System Components for User Story 1

User story 2: As a [data provider], I want [to publish] a [stream of time series

data] so that I can promote use of my sensor data, meet journal / funding requirements, and gain recognition for data outputs

Description: A [data provider] wants to publish and share a stream of data (from a sensor) with the user community for ongoing analysis. This data service needs to be citable in order to create credit for the data provider. He/she goes to the LTER Europe node in order to document and upload the information and enable others to use and cite this service.

The data service needs to have:

1. a complete MD description using a defined standard used in the community to provide information on the type of service and the content and origin of the time series
2. a persistent identifier (PID) that can be dereferenced to a metadata data record landing page that allows access to the data service
3. a data access services so that other can access a defined slice of the data series

Candidate Tasks

1. Enable metadata authoring through the editing tools in GET-IT EDI and assign PID
2. Harvest metadata records into the DIP repository so that the service can be discovered
3. Enable search and view of the metadata record in the DIP data discovery layer
4. Enable visualisation of the time series data from the data service
5. Enable download of the required slice of data via the data service link

Acceptance Criteria

1. Add a metadata record of the required data service
2. Upload the metadata record to the PID so that it is accessible from the discovery tools
3. Use the PID as a dereferencable link to the metadata page
4. Visualise the time series through the data service
5. Select a slice of data and download it

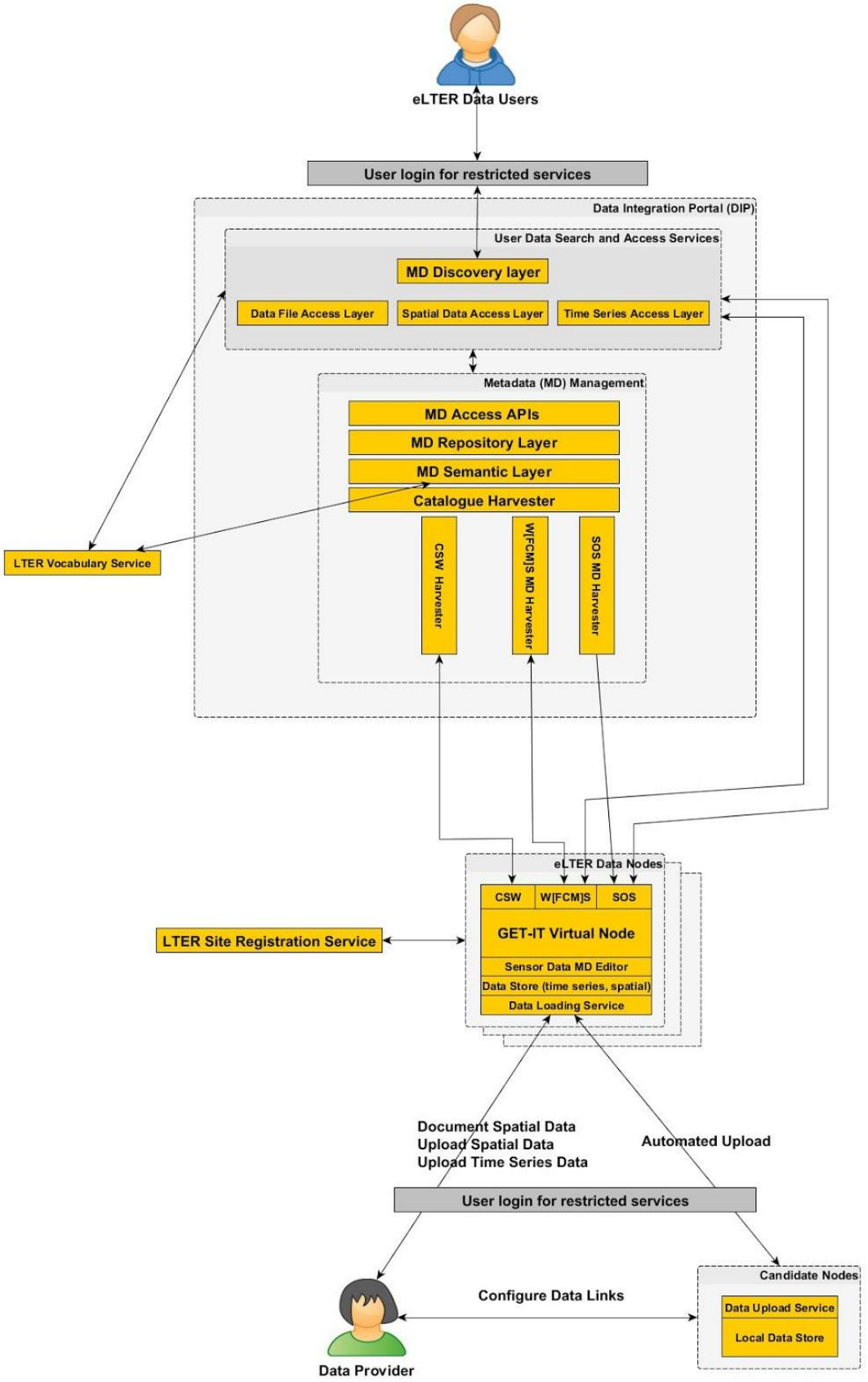

Figure 3: System Components for User Story 2

User story 3: As a [site network manager], I want [to join the network] with an

[existing data node] so that I can enable my site managers to publish their data to eLTER

Description: A [network manager] wants to join the existing eLTER data network to publish and share data files and data streams with the user community to publicise their site network data for reuse and citation. The data services need to be visible through the eLTER DIP and data accessible to eLTER users.

The network manager needs to have:

1. An administration service to document the details of the sites and data services to be added
2. A metadata harvesting service to pick up information from the data catalogues and data services
3. Services within the DIP to search for, find and access the data services provided by the new node.

Candidate Tasks

1. Enable metadata forms within DEIMS and/or DIP to allow new data services to be described and added
2. Harvest metadata records into the DIP repository so that the service can be discovered
3. Enable search and view of the metadata record in the DIP data discovery layer of the new data node
4. Enable access to data files available from the new data service

Acceptance Criteria

1. Configure metadata to add new node to DIP registry
2. Upload the metadata records to the PID so that it is accessible from the discovery tools
3. Enable access to data view and download services from the new data node
4. Visualise the time series through the data service
5. Download file and time series data

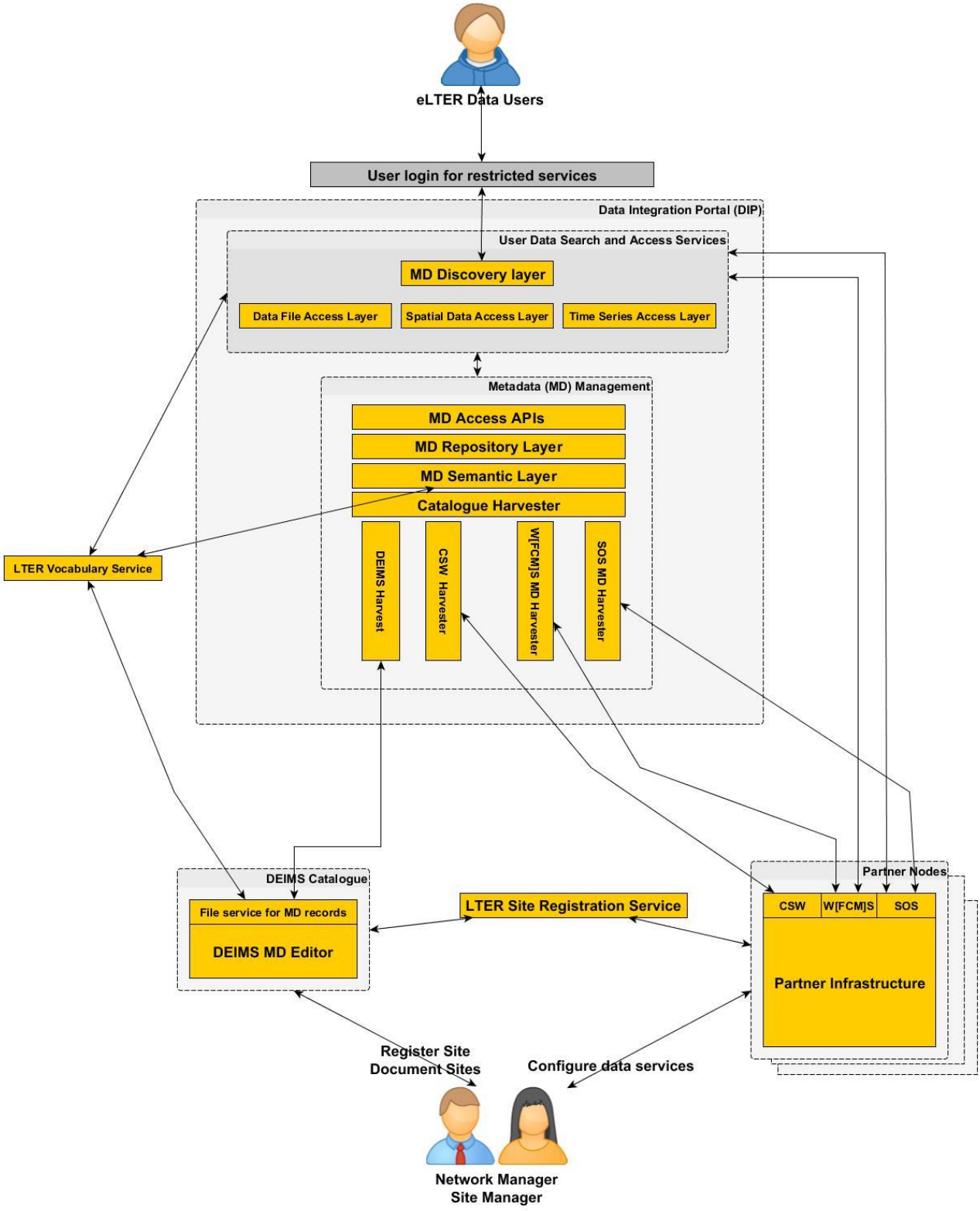

Figure 4: System Components for User Story 3

4 Glossary

Use this for acronyms mainly

Term	Definition
API	An application programming interface (API) is a set of subroutine definitions, protocols, and tools for building application software. In general terms, it's a set of clearly defined methods of communication between various software components
CSW	Catalog Service for the Web (CSW), sometimes seen as Catalog Service - Web, is a standard for exposing a catalogue of geospatial records in XML on the Internet (over HTTP). The catalogue is made up of records that describe geospatial data (e.g. KML), geospatial services (e.g. WMS), and related resources
DEIMS	The Dynamic Ecological Information System providing a site and dataset catalogue enabling to author, discover and share metadata on long term research sites, datasets, data products, persons, and networks.
DIP	The eLTER Information System Data Integration Portal
EUDAT	The EUDAT Collaborative Data Infrastructure is essentially a European e-infrastructure of integrated data services and resources to support research.
GET-IT	GET-IT Geoinformation Enabling ToolkIT starterkit®, is the open-source software suite developed by researchers from IREA and ISMAR, two Institutes of the National Council of Research (CNR) of Italy, in the context of the RITMARE Flagship Project and LifeWatch Italy . The suite is the first open-source collaborative effort toward the integration of traditional geographic information with observational data. This is achieved by coupling GeoNode with the SOS implementation by 52ºNorth and with components developed from scratch in the context of RITMARE, primarily for addressing data visualisation and metadata editing.
ISO	The International Organization for Standardization (ISO) is an international standard-setting body composed of representatives from various national standards organizations.
LDAP	Lightweight Directory Access Protocol (LDAP) is a client/server protocol used to access and manage directory information. It reads and edits directories over IP networks and runs directly over TCP/IP using simple string formats for data transfer.
MD modules	MetaData modules within the DIP that manage, transform and deliver metadata to users
OGC	The Open Geospatial Consortium (OGC), an international voluntary consensus standards organization, originated in 1994. In the OGC, more than 500 commercial, governmental, nonprofit and research organizations worldwide collaborate in a consensus process encouraging development and implementation of open standards for geospatial content and services, sensor web and Internet of

	Things, GIS data processing and data sharing.
RITMARE	RITMARE Flagship Project is one of the National Research Programmes funded by the Italian Ministry of University and Research. RITMARE is the leading national marine research project for the period 2012-2016; the overall project budget amount to 250 million euros, co-funded by public and private resources.
SOS	The Sensor Observation Service (SOS) is a web service to query real-time sensor data and sensor data time series and is part of the Sensor Web
W[FCM]S	The Open Geospatial Consortium Web Feature Service Interface Standard (WFS) provides an interface allowing requests for geographical features across the web using platform-independent calls. The Open Geospatial Consortium Web Coverage Service Interface Standard (WCS) defines Web-based retrieval of coverages – that is, digital geospatial information representing space/time-varying phenomena A Web Map Service (WMS) is a standard protocol for serving (over the Internet) georeferenced map images which a map server generates using data from a GIS database.