## DELIVERABLE D4.1

### COMSODE publication platform

**Open Data Node - for test**

<table>
<thead>
<tr>
<th>Project</th>
<th>Components Supporting the Open Data Exploitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acronym</td>
<td>COMSODE</td>
</tr>
<tr>
<td>Contract Number</td>
<td>FP7-ICT-611358</td>
</tr>
<tr>
<td>Start date of the project</td>
<td>1st October 2013</td>
</tr>
<tr>
<td>Duration</td>
<td>24 months, until 31st September 2015</td>
</tr>
<tr>
<td>Date of preparation</td>
<td>3.10.2014</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Peter Hanečák, Svetozár Krchnavý, Gabriel Lachmann, Tomáš Knap</td>
</tr>
<tr>
<td>Responsible of the deliverable</td>
<td>Peter Hanečák</td>
</tr>
<tr>
<td>Email</td>
<td><a href="mailto:peter.hanecak@eea.sk">peter.hanecak@eea.sk</a></td>
</tr>
<tr>
<td>Reviewed by</td>
<td>Ivan Hanzlík</td>
</tr>
<tr>
<td>Status of the Document</td>
<td>final version</td>
</tr>
<tr>
<td>Version</td>
<td>1.0</td>
</tr>
<tr>
<td>Dissemination level</td>
<td>PU (Public)</td>
</tr>
</tbody>
</table>
Table of Contents

1. Executive summary ........................................................................................................ 3
2. Deliverable context ......................................................................................................... 3
   Purpose of the deliverable .............................................................................................. 3
   Related documents ....................................................................................................... 3
3. Methodology used ......................................................................................................... 4
   Step 1 - development plan ............................................................................................ 4
   Step 2 - implementation of ETL & enrichment functionality ...................................... 6
   Step 3 - implementation of Publication functionality ................................................... 7
   Step 4 - implementation of Cataloguing functionality .................................................. 7
   Step 5 - implementation of Management functionality ................................................. 8
5. Test release of Open Data Node ..................................................................................... 9
   UnifiedViews ................................................................................................................. 9
   Virtuoso (Open Source version) .................................................................................... 10
6. Support infrastructure created for Open Data Node components ............................... 11
   Support infrastructure for Open Data Node itself ....................................................... 11
7. Provided APIs and data dumps ...................................................................................... 12
1. Executive summary

Deliverable 4.1 consists mainly of test release of COMSODE Open Data publication platform named Open Data Node (also referred to as ODN), along with the documentation sufficient to deploy, use and maintain the solution. The delivery is licensed under a set of Open Source licenses and all parts are available on the project website for free download.

Deliverable is a snapshot of what is publicly available as output of Task 4.1 at the end of M12 (Sep 2014, i.e. end of Year 1, major project milestone with several deliverables due). Task started in M8 and continues until M21 thus the output at M12 is partial, suitable only for testing purposes.

This document describes the main achievements of Task 4.1 available at the end of M12:
- test release of Open Data Node
- support infrastructure created for its components
- APIs and data dumps provided for application development

It also describes purpose of this deliverable and methodology used to produce above mentioned milestones.

2. Deliverable context

Purpose of the deliverable
The purpose of this deliverable is to build publicly available Open Source implementation of early-stage demonstrator for Open Data publication platform.

This early stage is then further used:
- to validate current approach to achieving project goals,
- to develop Deliverable 4.2 (Tailored search application on-top of data published using ODN),
- to achieve project indicators due in Year 1 and
- to support further Exploitation.

Related documents
List of related documents from project:
- D2.1 – User requirements for the publication platform from target organizations, including the map of typical environments
- D2.3 – Architecture and design documentation for COMSODE development tasks
- D5.1 – Methodology for publishing datasets as open data
- D4.2 - Tailored search application on-top of data published using ODN - for test
- D7.3 - Half-time exploitation strategy
- Periodic Progress Report, due November 2014
3. Methodology used

The final goal of Task 4.1 is a delivery of Open Source publication platform for Open Data (also referred to as Open Data Node, ODN in short) which is expected to be deployed mainly within data publishing organizations and which will provide access to published Open Data sets to general public.

Development process is part of Task 4.1 and started in M8 (May 2014) with inputs from WP2 (Deliverable 2.3). Task ends in M21 (June 2014).

Development is divided into 5 main steps:
1. creation of development plan
2. implementation of ETL & enrichment functionality
3. implementation of Publication functionality
4. implementation of Cataloguing functionality
5. implementation of Management functionality

At the end of project’s Year 1 (i.e. M12, Sep 2014), only steps number 1 and 2 are complete. Thus Deliverable 4.1 - test version of Open Data Node - consists of results from those two steps and is not feature complete. But the test version is suitable for achieving some project indicators (publication of subset of COSMODE datasets - see deliverables from WP3, for development of test version of Search application - D4.2 - and for further Exploitation).

Subsequent development steps will then further enhance the test version, utilizing further feedback from users of test version of ODN. The end result will be Deliverable 4.3: final version of Open Data Node.

Step 1 - development plan

Based on inputs from WP2 - for example high-level architecture of Open Data Node (Fig. 2) and effort estimates for each particular ODN module - it was decided to align the development roughly along the data flow paths: from the raw data and metadata at the beginning to the final Open Data (datasets and their metadata) at the end. Processes and steps described in Deliverable 5.1 (generic Open Data publishing methodology) follow similar pattern thus further reinforcing this decision.

So in Task 4.1, we’re developing Open Data Node roughly from “the left” to “the right” (Fig. 1):
Figure 1: Basic Open Data Node use-case, with flow of data illustrated

Figure 2: Architecture of Open Data Node (from D2.3) with implementations order included

Thus, the order of implementation was decided to be:
1. module ODN/UnifiedViews: ETL & enrichment functionality
2. module ODN/Storage: storage of results from ODN/UnifiedViews for later use in ODN/Publication
3. module ODN/Publication: functionality for publication of data dumps and APIs
4. module ODN/InternalCatalog: internal cataloguing functionality needed mainly for first phases of COSMODE Methodology (D5.1)
5. module ODN/Catalog: public cataloguing functionality needed mainly for final phases of COSMODE Methodology (D5.1)
6. module ODN/Management: management and integration functionality for Open Data Node as a whole

As some modules are more closely related to each other, final plan consists of 4 further steps described in subsequent chapters. Each step was estimated to take approx. 3 to 4 months to complete and steps are ordered one after another with small (one or two weeks) overlaps. **Figure 3 illustrates the roadmap.**

**Step 2 - implementation of ETL & enrichment functionality**

In this step, ODN development started by implementing ODN/UnifiedViews and (part of) ODN/Storage and ODN/Publication modules. Following tools were identified in WP2 for reuse:

- ODCS ([https://github.com/mff-uk/ODCS](https://github.com/mff-uk/ODCS)) as base for ODN/UnifiedViews module (originally named as ODN/ETL) with non-trivial development effort needed to enhance the tool to meet the needs of COMSODE project
- Apache HTTP project ([http://httpd.apache.org/](http://httpd.apache.org/)) as one of components for ODN/Storage module, reused as-is, without need to modify it
- Virtuoso - Open Source version ([http://virtuoso.openlinksw.com/](http://virtuoso.openlinksw.com/)) as another component for ODN/Storage module, reused as-is, without need to modify it

**Figure 3 (right): Open Data Node development roadmap, showing also relation to Task 4.3, Task 6.5, inputs from other Work Packages and also Year 1 milestone (due date for this Deliverable 4.1)**
Majority of development effort went into redesign and stabilization of ODCS, which was also - as part of this step (and with collaboration with associated partners) - rebranded to UnifiedViews (see the note below). From the point of view of this deliverable, this step provided sufficient outputs for test version of Open Data Node.

Note: Some part of development effort (partially also Exploitation) related to UnifiedViews actually started in M6 (March 2014), well ahead of planned start of Step 2 (M8). This was needed to smoothly transition from ODCS student project into community backed UnifiedViews. Initial community consisting of EEA (COMSODE consortium partner), Semantic Web Company (COSMODE consortium associated partner) and Semantica.cz (spin-off company of some of original ODCS developers) was formed, governance structure created and licensing and development model put in place. In line with that, also module ODN/ETL was renamed to ODN/UnifiedViews so as to signify that this module is tasked not just with ETL operations but also with data enhancement and enrichment functions, as described in DoW.

**Step 3 - implementation of Publication functionality**
ODN development continues in this step mainly with implementation of ODN/Publication module, with further enhancements to ODN/UnifiedViews and ODN/Storage.

WP2 identified tools for reuse, for example:

- restSQL ([http://restsql.org/](http://restsql.org/)) as base for implementation of REST API for datasets published via Open Data Node, non-trivial development effort dedicated to enhance the tool to meet the needs of COMSODE project
- Virtuoso - Open Source version ([http://virtuoso.openlinksw.com/](http://virtuoso.openlinksw.com/)) as provider of SPARQL endpoint functionality, to be reused as-is, without need to modify it

From the point of view of this deliverable, this step is work in progress and did not yet provide sufficient outputs for test version of Open Data Node.

**Step 4 - implementation of Cataloguing functionality**
ODN development in this step consists mainly of implementation of two modules: ODN/InternalCatalog and ODN/Catalog. Implementation was merged into one step because WP2 identified same tools to be reused for both and also because both modules are closely related through data publication use-cases.

WP2 identified following tool for reuse in both modules:

- CKAN ([http://ckan.org/](http://ckan.org/)) as base for cataloguing functionality in Open Data Node, non-trivial development effort is expected to enhance the tool to meet the needs of COMSODE project
In each module, CKAN will require different set and amount of customization, majority of it in ODN/InternalCatalog.

As this step has not started yet, from the point of view of this deliverable it did not yet provide any outputs for test version of Open Data Node.

**Step 5 - implementation of Management functionality**

ODN development in this step consists mainly of implementation of module ODN/Management. WP2 has not identified any major tool for reuse in this modules as majority of the work is expected to be customization and integration between tools reused in other modules.

As this step has not started yet, from the point of view of this deliverable it did not yet provide any outputs for test version of Open Data Node.
5. Test release of Open Data Node

Test version of Open Data Node consists of module ODN/UnifiedViews implemented using UnifiedViews.

This is supplemented with preliminary version of following two modules:

1. ODN/Storage - implemented using general file system and Virtuoso (Open Source version)
2. ODN/Publication - very early implementation based on Apache HTTPD

Main COMSODE development effort was dedicated to UnifiedViews so this will be described in more details. Some notes will be included also about Virtuoso, as only its certain versions are supported.

UnifiedViews

COMSODE project contributed and released so far 8 versions of UnifiedViews, starting with version 0.9 (in March 2014) and currently at version 1.3.1 (in September 2014).

UnifiedViews is available under set of Open Source licenses and can be obtained:

- as source code from [https://github.com/UnifiedViews](https://github.com/UnifiedViews)

UnifiedViews consists of several components:

- Core: Main part, providing management GUI and backend running ETL and enrichment task (called pipe lines), which are constructed from individual Data Processing Units (DPUs).
  - [https://github.com/UnifiedViews/Core](https://github.com/UnifiedViews/Core)
  - license: combination of GPLv3 and LGPLv3, commercial licensing can be obtained
- Plugins: Set of common DPUs provided for Core.
  - [https://github.com/UnifiedViews/Plugins](https://github.com/UnifiedViews/Plugins)
  - license: LGPLv3, commercial licensing can be obtained
- Plugin Development Environment: Tools and libraries needed by programmers for development of new or modification of existing DPUs.
  - [https://github.com/UnifiedViews/Plugin-DevEnv](https://github.com/UnifiedViews/Plugin-DevEnv)
  - license: combination of GPLv3 and LGPLv3, commercial licensing can be obtained
- Packages: Sources needed to build UnifiedView packages for various Linux distributions.
  - [https://github.com/UnifiedViews/Packages](https://github.com/UnifiedViews/Packages)
  - license: combination of GPLv3, LGPLv3 and Public Domain, commercial licensing can be obtained
Figure 4: Screenshot of UnifiedViews administration GUI showing list of data processing pipelines.

Figure 5: Screenshot of UnifiedViews showing detailed configuration of one particular data processing pipeline, which consists of several DPUs.

**Virtuoso (Open Source version)**

COSMODE project reuses Open Source version of Virtuoso as is, without any modifications or customizations. Virtuoso versions 6.1.x and 7.0.x are supported by Open Data Node/UnifiedViews.

Virtuoso Open Source edition can be obtained from [http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSDownload](http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSDownload) as source code or pre-built binaries for several platforms.
6. Support infrastructure created for Open Data Node components

Support infrastructure for Open Data Node itself
COMSODE prepared following infrastructure for Open Data Node:
  ● source code repository: https://github.com/OpenDataNode/open-data-node
  ● public Wiki: https://utopia.sk/wiki/display/ODN/Open+Data+Node+Home

As the test version of Open Data Node is based mainly on UnifiedViews, this infrastructure is not used much for now.

Support infrastructure for UnifiedViews
COMSODE contributed to preparation of following infrastructure for UnifiedViews:
  ● public web site: http://www.unifiedviews.eu/
  ● source code repository: https://github.com/UnifiedViews
  ● public issue tracking system: https://github.com/UnifiedViews/Core/issues
  ● public Wiki: https://grips.semantic-web.at/display/UDDOC/Introduction
    ○ Administrator and Developer Guide: https://grips.semantic-web.at/display/UDDOC/UnifiedViews+Administrator+and+Developer+Guide
  ● mailing lists:
    ○ public: for broader community support: unifiedviews@googlegroups.com
    ○ private: for development collaboration: unifiedviews-dev@googlegroups.com
7. Provided APIs and data dumps

As part of Task 6.3 (Pilot publishing project) test version of Open Data Node is deployed on COMSODE infrastructure and used by the consortium to publish (as of this date) 42\(^1\) COMSODE datasets.

For those 42 datasets, test instance of Open Data Node provides:

   ○ note: This is a preliminary functionality, given that it is a scope of Step 4, which has not started yet (see page 7).
   ○ note: RDF format in Turtle notation is used and ZIP compression is employed to reduce storage and connectivity requirements.
   ○ note: This is a preliminary functionality, given that it is a scope of Step 3, which is still work in progress (see page 7).

Data published this way are also used in test version of Tailored search application delivered by consortium and described in Deliverable 4.2.

This service, along with test Search application, are now providing real-world feedback for further ODN development effort.

Note: REST APIs are in the scope of Step 3 (see page 7) and fully functional public data catalog is in the scope of Step 4 (see page 7) so those are not part of this deliverable.

\(^1\) Project is releasing new datasets continually, so the number of datasets reported on this catalog is increasing over time.