EUCLID

EdUcational Curriculum for the usage of LInked Data

Deliverable 1.1.3

Curriculum delivery

Coordinator: Maria Maleshkova

With contributions from: Maribel Acosta, Alexander Mikroyannidis

Quality Assessor: Elena Simperl

Editor:	Maria Maleshkova, KIT
Deliverable nature:	Other (O)
Dissemination level: (Confidentiality)	Public (PU)
Contractual delivery date:	31.10.2012
Actual delivery date:	31.10.2012
Version:	1.0
Total number of pages:	42
Keywords:	Curriculum plan, training modules, learning syllabus

Executive summary

This deliverable describes the final version of Euclid's curriculum, including details on the content of each of the training modules, the available materials, and its alignment with other existing training programmes and initiatives.

Current developments on the Web are marked by the increasing importance and application of Linked Data technologies, which are establishing themselves as an innovative way for publishing, interlinking and exploring data sets. Until now the main adopters have been research organisations, governmental institutions and a selection of companies. However, with the increasing number of interested parties committing to use Linked Data core principles, and through the plenitude of applications build on top of the already available data, it seems that the time is right for the wider industrial adoption of Linked Data. EUCLID aims to support this development by addressing the need for trained data practitioners who are able to apply a Linked Data approach as part of their data business solutions.

In particular, one of the main objectives of the project is to provide an extensive training curriculum, covering the main technologies, tools, use cases and competencies that need to be acquired in order to complete both basic, as well as more complex, tasks related to practically using Linked Data. This deliverable describes the final version of the curriculum, including the adjusted content of the individual modules, the different learning methods and materials, the targeted skills level that is to be achieved, and the expected competencies. In particular, the curriculum has been updated both in terms of the overall structure of the modules, as well as regarding the content of the individual modules themselves.

The second part of the deliverable details the curriculum plan. This is the final version of the curriculum, which includes final updates of the modules, rescoping of the content and improving the examples, tools and exercises. The curriculum presented here is divided into six modules, covering three levels of topics – introductory, advanced and expert. We continue to follow the approach of having modules build on each other, including more specific knowledge such as visualisation approaches or building applications on top of Linked Data. The redefined six modules are – 'Introduction to Linked Data and Application Scenarios', 'Querying Linked Data', 'Providing Linked Data', 'Interaction with Linked Data', 'Creating Linked Data Applications', and 'Scaling-up'. The past six months, since D1.1.2, were devoted to releasing the content on 'Interaction with Linked Data', preparing the materials for the module on 'Creating Linked Data Applications' and determining the relevant topics and structure of the final chapter on 'Scaling-up'. Furthermore, the scope of some of the individual modules was also continuously improved and redefined, since the main aim is to support data practitioners in completing different tasks and not simply to teach certain technologies or tools. Furthermore, improvements were done based on feedback and comments, gathered after sharing and presenting the curriculum. However, the main driving factor for both the overall content of the curriculum and the subsequently implemented refinements was the practical orientation of the content, examples, and exercises in each module, and of the curriculum as a whole.

The third part of this deliverable introduces a number of related programs and initiatives, which aim to provide training on Linked Data and data management and analysis in general. This section contains more details and is further refined in comparison to the first version of D1.1.2. In particular, we aim to identify existing courses and curricula, determine the main topics that they address and identify overlaps with the EUCLID curriculum. We show how they can be used in combination with EUCLID's modules in order to acquire competencies in cross-disciplinary fields in the general area of data science. This alignment of EUCLID's curriculum with further training activities and curricula is an important contribution of the deliverable, since it demonstrates the direct applicability of the produced learning materials in ordered to train a wide-range of data practitioners.

The final section of the deliverable includes a summary of related training events, courses and organisation. This information helps us not only to determine, which are currently the most popular topics but also helps us to identify the main players, which might also turn out to be prospective collaboration partners. Furthermore, we can track their activities and keep the here presented curriculum up-to-date and aligned with commonly offered training events and courses.

Document Information

IST Project Number	FP7 - 296229	**Acronym**	EUCLID
Full Title	Educational curriculum for the usage of Linked Data		
Project URL	http://www.euclid-project.eu/		
Document URL	http://www.euclid-project.eu/resources/deliverables/		
EU Project Officer	Martina Eydner		

Deliverable	**Number**	1.1.3	**Title**	Curriculum Delivery
Work Package	**Number**	1	**Title**	Course Production and Delivery

Date of Delivery	**Contractual**	M18	**Actual**	M18
Status	version 1.0		final - ☑ ☐	
Nature	other			
Dissemination level	public			

Authors (Partner)	Maria Maleshkova (KIT)			
Responsible Author	**Name**	Maria Maleshkova	**E-mail**	maria.maleshkova@kit.edu
	Partner	KIT	**Phone**	+49 721 608 4 7363

Abstract (for dissemination)	One of the main objectives of EUCLID is to provide and extensive training curriculum, covering the main technologies, tools, use cases and skills that need to be acquired in order to complete both basic, as well as more complex, tasks related to dealing with Linked Data. This deliverable describes the final version of the curriculum, including the adjusted content of the individual modules, the different learning methods and materials, and the expected competencies.
Keywords	Curriculum plan, training modules, training syllabus

Version Log			
Issue Date	**Rev. No.**	**Author**	**Change**
27.05.2013	0.1	Maria Maleshkova (KIT)	Initial deliverable draft
09.09.2013	0.2	Maria Maleshkova (KIT)	Refined section on curriculum alignment and further training initiatives
25.09.2013	0.3	Maria Maleshkova (KIT)	Improved sections
06.10.2013	0.4	Maria Maleshkova (KIT)	First draft submitted for internal review
28.10.2013	1.0	Maria Maleshkova (KIT)	Final draft

Table of Contents

List of Figures and Tables

Abbreviations

DBMSs – Database Management System standard

DL – Distance Learning

DoW – Description of Work

EU – European Union

IT – Information Technology

KR – Knowledge Representation

LD - Linked Data

LOD – Lined Open Data

OWL – Ontology Web Language

OWL-S – OWL for Services/ OWL-based Web Service Ontology (formerly DAML-S)

RDF/S – Resource Description Framework / Schema

RDBMS - Relational Database Management System

SPARQL – SPARQL Protocol and RDF Query Language

SQL – Structured Query Language

WP – Work Package

WSMO – Web Service Modelling Ontology

XML - Extensible Markup Language

1 Introduction

Current developments in the context of sharing data on the Web are marked by the growing importance and use of Linked Data, which is becoming a de-facto standard to publish and access structured data on the Web. This trend is supported by the increased number of governmental organizations, research institutes, but also companies, which deal with data exchange, manipulation and maintenance in their daily business activities. As a result, there is a need for trained data practitioners, who can apply Linked Data solutions in different contexts and as part of various solutions. EUCLID aims to address precisely this demand by providing an extensive training curriculum, backed up by a range of social and community-based activities, which aim to disseminate but also to gather feedback about the provided training materials.

This deliverable presents the final version of the Euclid's training curriculum, which includes final updates of the modules, rescoping of the content and improving the examples, tools and exercises. The curriculum described here is divided into six modules, covering three levels of topics – introductory, advanced and expert. We continue to follow the approach of having modules build on each other, including more specific knowledge such as visualisation approaches or building applications on top of Linked Data. The redefined six modules are – 'Introduction to Linked Data and Application Scenarios', 'Querying Linked Data', 'Providing Linked Data', 'Interaction with Linked Data', 'Creating Linked Data Applications', and 'Scaling-up'. The past six months, since D1.1.2, were devoted to releasing the content on 'Interaction with Linked Data', preparing the materials for the module on 'Creating Linked Data Applications' and determining the relevant topics and structure of the final chapter on 'Scaling-up'. Furthermore, the scope of some of the individual modules was also continuously improved and redefined, since the main aim is to support data practitioners in completing different tasks and not simply to teach certain technologies or tools. Furthermore, improvements were done based on feedback and comments, gathered after sharing and presenting the curriculum. However, the main driving factor for both the overall content of the curriculum and the subsequently implemented refinements was the practical orientation of the content, examples, and exercises in each module, and of the curriculum as a whole.

In summary, the curriculum plan takes a practice- and application-oriented approach towards communicating essential Linked Data knowledge that would help data practitioners to apply this new technology in the context of their daily tasks.

In the following sections we briefly finalise the definition of our targeted audience, the goal level of knowledge, the skills and knowledge prerequisites, the delivery methods and the materials. These characteristics were discussed in more detail in the first two versions of the deliverable. Therefore, here, we only state out concluding findings.

1.1 Target Audience and Targeted Level of Knowledge

The main target audience of EUCLID's curriculum are data practitioners and professionals, who already use or aim to adopt Linked Data as means to publish and access structured data over the Web. This also motivates the practical orientation of the modules and the use of directly appreciable examples.

However, the experience gathered throughout the project duration has clearly demonstrated that, in fact, the materials are used by a very broad audience, including researchers, students, professionals, managers, technology experts, ect. Therefore, the curriculum and the trainings can be of benefit for anyone who aims to gain a broader and deeper understanding of how to manage data in accordance with Linked Data principles.

In terms of the targeted skills and knowledge that are to be gained, the curriculum provides three main levels of expertise:

- Introductory Level – This level is based on a set of modules that communicate the fundamental skills that are required in order to begin applying Linked Data technologies.
- Basic Level – The second level deals with more advanced topics, also specializing in different areas such as visualization and query processing.
- Advanced Level – This level aims to provide expertise knowledge that is rather specific to an area of use and requires somewhat extensive prior knowledge.

By defining the modules, and the curriculum as a whole, in terms of such a progressive structure, it is not

necessary to try to bring all trainees to the same level. Instead, the learning plan can be adapted in order to address the particular needs of the audience and serve both professionals, who strive to become experts or want to simply get some basic knowledge in the area, alike.

1.2 Skills and Knowledge Prerequisites

In order to be able to grasp the main concepts, the application functions and the presented approaches, some previous knowledge in IT development and engineering are very useful. The lack of experience in a particular area can be compensated for by giving examples and step-by-set guides, which demonstrate how the learned principles and techniques can be applied. The more advanced modules can benefit from some knowledge in the corresponding fields.

1.3 Delivery Methods and Materials

In this section we present the final set of delivery methods and materials, which were developed based on the experience gained within the project. Overall, the curriculum plan is based on a set of delivery methods and materials that complement each other and updates of the curriculum influence both the format and the ways of communicating the content.

1.3.1 Delivery Methods

We employ two main methods for delivering the learning content – via online channels and directly by a professional trainer. Given that the main target group of the curriculum plan is data professionals, self-training and distance learning remain the main means of communicating the courses content. These types of training methods are relatively flexible when it comes to geographical location and time-slot allocation and are, therefore, suitable for on-the-job, but also parallel-to-the-job training.

Online communication channels, such as platforms for sharing slides, videos or complete training courses are very useful for supporting self-training and distance learning. In fact, we found out that SlideShare[1] proves to be extremely applicable in terms of sharing and disseminating the project results. The achieved outreach is much greater and really anyone interested in the topic can benefit. Furthermore, the online webinars were very successful, enabling a high number of simultaneous views and overcoming geographical boundaries. More details on the online communication channels will be provided in D2.1.5 “Final online community engagement report”.

In contrast to self-training based on online resources, distance-learning provides more guidance to the students in terms of the learning plan but also the support in terms of interaction with trainers or gathering feedback. It is not within the scope of EUCLID to organise a distance-learning event, however, the materials are well suited to be used as a basis for such a course. In particular, the combination of the live webinars, guided tutorials, and the official course materials can easily support such a learning approach.

EUCLID’s materials have also proven to be very useful as part of on-site trainings. The trainings were conducted both for professionals, as well as for people with more of a research background. Further details on the training events, the corresponding topics, audience and expected impact are available in Deliverable 2.3.3 “Final real-world community engagement report”. For example, the sources of the ESWC Summer School 2013[2] were created based on EUCLID materials and were very well received by the participating PhD students.

1.3.2 Training Materials

In this section we present the final list of formats, in which we offer the EUCLID training materials. Both the content-generation process, as well as the format of the training materials themselves, have undergone improvement and reshaping.

[1] http://www.slideshare.net/EUCLIDproject
[2] http://summerschool2013.eswc-conferences.org/

EUCLID generates the following materials for each of the curriculum's modules:

- **Description and detailed outline of the module**

The description of the module as well as the detailed outline serve as a starting point for specifying the content, the technologies, tools and examples that are to be covered by the module. These represent a short and concise way of introducing the covered topics and providing an overview of the content. This type of content is reused on the website, for describing presentations on SlideShare and for colleting feedback.

- **Related materials and further reading**

These materials provide pointers to relevant presentations, videos, books and articles that can be used to deepen the acquired knowledge. We also include sources that were used as a basis for the content of some of the chapters, since the trainees should be aware of the standard literature in the field.

- **Presentations on specific topics**

During the course of developing modules 3 and 4, it became evident that it would be very helpful to provide training on some related topics, which are either more advanced or represent an emerging trend, and cannot be directly included as part of the modules. Therefore, we also provide some additional presentations on further topics, such as Relational Database to RDF (RDB2RDF) as an annex to the module on Providing Linked Data. Furthermore, we will continuously include additional tools to the already completed modules and add a general ontology engineering training.

- **Examples, exercises and multiple-choice questions**

The examples are included as part of the slides and the written chapter, in order to directly demonstrate how the learned theory can be used in practice. In fact, some of the topics can be effectively conveyed only with the help of good examples, such as for instance "Querying Linked Data with SPARQL". The exercises are usually at the end of each module and are part only of the complete multimedia versions of the content. However, the examples that occur in the slides, are discussed in the webinars and are naturally a part of the website and the textual materials. Each module is concluded with a set of multiple-choice questions, which are an effective means for self-assessment, requiring to apply some of the learned principles or to use introduced tools.

- **Presentation slides**

Each module has a set of PowerPoint slides, which can be used for doing trainings. The presentations slides can be rearranged and combined in order to deliver trainings to a particular topic or for a specific target audience. The slides cover the module-topics in detail and are available both on the website and on EUCLID's SlideShare channel.

- **Webinars**

The webinars are conducted based on the slides for each module. The webinars are broadcasted live from the Podium facility of the Open University[3], after which a recording is made available both from the same facility and also through the EUCLID channel in Vimeo[4]. Furthermore, parts of the webinars are integrated directly in the versions of the module available as eBooks.

- **Screencasts of tools**

The EUCLID screencasts consist of short clips (2-3 minutes) that provide a quick overview and a walkthrough of a representative set of tools and platforms related to the EUCLID modules. The screencasts are made available in the EUCLID Vimeo channel and are also included in the eBook chapters in order to better explain the relevant sections of the chapters.

- **eBooks**

[3] http://stadium.open.ac.uk/podium
[4] https://vimeo.com/euclidproject

The EUCLID's eBooks have a multimedia format, including detailed text description, embedded videos, images, exercises and they encompass all the content for each module in a structured and interactive way. The eBook serves as the basis for self-learning, as well as for revisiting certain topics after a training is completed. In order to maximise the impact of the training materials on the community and bring them closer to as many people as possible, the eBook chapters are made available for a variety of platforms and formats including:

- Web browsers (HTML format)
- Apple iPad (iBook format)
- eReaders (ePUB format)
- Amazon Kindle devices (AZW3 format)

In summary, the eBooks represent the final outcome of the training materials revising process. All the modules will be published in iTunesU[5] and will be freely available. In this way EUCLID materials will be published in one of the most popular platforms for online learning and can be used individually or in conjunction with further courses.

1.3.3 Finalised Materials Production Process

One of the main lessons learned during the course of the project is that the collaborative creation of learning content requires a well-specified and structured materials production process. See Figure 1 visualised the final version of the applied production process, which underwent a number of revisions and improvements.

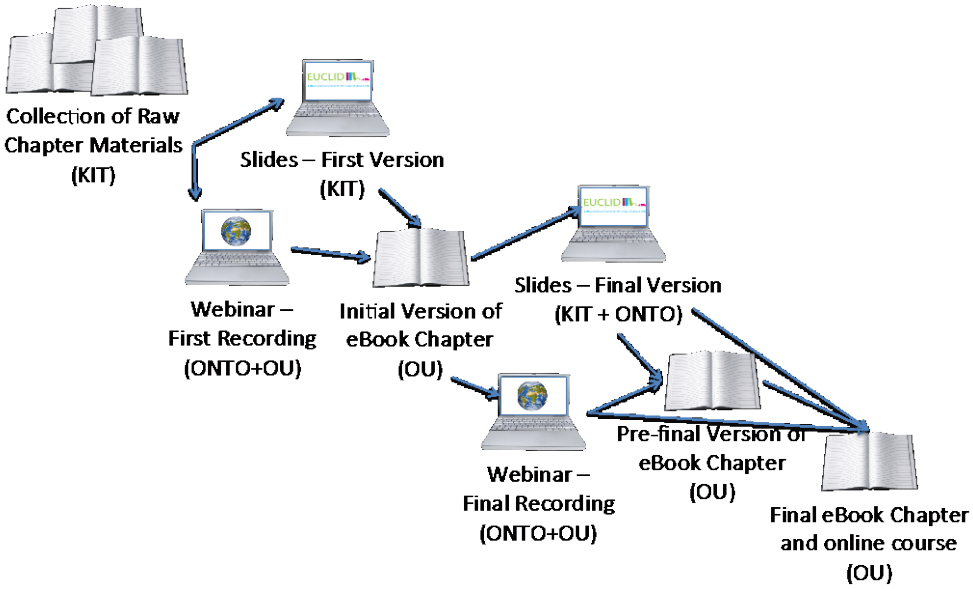

Figure 1: Final Materials Production Process

The steps of the final materials production process are the following:

1. Defining the detailed outline of the chapter and defining the content to be covered
2. Collect related materials, announce the dates for the webinars and the release
3. Discuss and finalise the proposed content
4. Create finalised outline, exercises and examples
5. First version of the slides
6. Internal webinar

[5] http://www.apple.com/apps/itunes-u/

7. Second version of the slides, first version of the HTML chapter, internal quality assurance
8. Finalise exercises and examples
9. Finalise slides, finalise HTML chapter
10. Public webinar
11. Public release

Overall, EUCLID implements an approach of revising the training content based on gathered feedback. For example, after the initial webinar, comments and suggestions are gathered from the audience. Furthermore, each draft of the modules is made available online and community feedback is actively be gathered. As a result, the final version of the training materials has undergone at least two rounds of improvements and revision.

2 Finalised Structure of the Curriculum

This section describes the finalised curriculum, including details on the encompassed six modules.

Figure 2: Curriculum Structure

As previously reported, the EUCLID's curriculum covers one introductory module, three basic modules, and two advanced modules (see Figure 2). This distribution of the modules is the result of a number of revisions and rescoping of the initial curriculum plan. The following sections describe each of the modules in more detail, providing a summary of the content, detailed outline of the topic, and a list of the already available materials.

Please refer to http://www.euclid-project.eu/resources/learning-materials for the latest pointers and URL of current materials. Some of the URL in this document might become outdates, since the iBooks are about to be published on the iTunes store and the different formats of the written modules will be available through the same link, which enables the downloading of the content in the appropriate device-dependent file type.

2.1 Introduction and Application Scenarios

This module introduces the main principles of Linked Data, the underlying technologies and background standards. It provides basic knowledge for how data can be published over the Web, how it can be queried, and what are the possible use cases and benefits. As an example, we use the development of a music portal (based on the MusicBrainz[6] data set), which facilitates access to a wide range of information and multimedia resources relating to music. The module also includes some multiple-choice questions in the form of a quiz, screencasts of popular tools and embedded videos.

Following is a detailed outline of the module as well as a summary of all the available materials.

2.1.1 Module Detailed Outline

All the materials to the first module are available at http://www.euclid-project.eu/#chapter1. The following listing describes the covered topics in detail.

1.1 Introduction

[6] http://musicbrainz.org

1.2 Motivation of the Course
1.3 Background Technologies
- Internet
- Hypertext
- WWW
- Web 1.0 (static)
- Web 2.0 (dynamic)
- Social Web
- Web 3.0 (semantic)
- Ontologies

1.4 Background Standards
- HTTP
- URI
- XML
- RDF
- RDFS
- OWL (OWL 2 Full, OWL 2 DL, OWL 2 EL, OWL 2 QL, OWL 2 RL)
- SPARQL

1.5 Linked Data
- Linked Data Principles
- Rating Published Datasets
- Growth of Linked Data on the Web

1.6 Case Scenario: a Music Portal
1.7 Examples
- Marbles
- Sigma
- DBpedia Mobile

2.1.2 Available Materials

The introductory module is the first module, for which all materials are already available. All the materials to module 1: Introduction and Application Scenarios are available at:

- Outline
  - http://www.euclid-project.eu/resources/curriculum
- Slides
  - http://www.slideshare.net/EUCLIDproject/usage-of-linked-data-introduction-and-application-scenarios
- Webinar and screencasts
  - Webinar Part I https://vimeo.com/61612182
  - Webinar Part II https://vimeo.com/61612378
  - Screencast: Sig.ma https://vimeo.com/57931687
  - Screencast: Data.gov.uk https://vimeo.com/57931688
  - Screencast: BBC Music https://vimeo.com/57931689
  - Screencast: MusicBrainz https://vimeo.com/57935375
- Exercises
  - Quiz http://www.euclid-project.eu/node/30/take

- eBook chapter
  - HTML http://www.euclid-project.eu/modules/chapter1
  - iBook http://www.euclid-project.eu/sites/default/files/resources/Chapter1.ibooks
  - ePUB http://www.euclid-project.eu/sites/default/files/resources/Chapter1.epub
  - Kindle http://www.euclid-project.eu/sites/default/files/resources/Chapter1.azw3

2.2 Querying Linked Data

Querying Linked Data is the first of the two advanced topic modules. This module looks in detail at SPARQL (SPARQL Protocol and RDF Query Language) and introduces approaches for querying and updating semantic data. It covers the SPARQL algebra, the SPARQL protocol, and provides examples for reasoning over Linked Data. The module uses examples from the music domain, which can be directly tried out and ran over the MusicBrainz dataset. This includes gaining some familiarity with the RDFS and OWL languages, which allow developers to formulate generic and conceptual knowledge that can be exploited by automatic reasoning services in order to enhance the power of querying.

2.2.1 Module Detailed Outline

All the materials to the second module are available at http://www.euclid-project.eu/#chapter2. The following listing describes the covered topics in detail.

2.1 Introduction and Motivation Scenario

2.2 SPARQL Terminology

2.3 Querying and Updating Linked Data with SPARQL

- Introduction to SPARQL
- Querying Linked Data with SPARQL
  - Query forms: ASK, SELECT, DESCRIBE, CONSTRUCT
  - Query patterns: BGP, UNION, OPTIONAL, FILTER
  - Sequence modifiers: DISTINCT, REDUCED, ORDER BY, LIMIT, OFFSET
- Updating Linked Data with SPARQL 1.1
  - Data management: INSERT, DELETE; DELETE/INSERT
  - Graph management: LOAD, CLEAR, CREATE, DROP, COPY/MOVE/ADD
- SPARQL Protocol: query operation, update operation

2.4 Reasoning over Linked Data

- SPARQL 1.1 entailment regimes
- RDFS entailment regimes, lacks of consistency check, inference limitations
- OWL properties, property axioms, axioms, class constructions

2.2.2 Available Materials

The currently available materials to module 2: Querying Linked Data can be found at:

- Outline
  - http://www.euclid-project.eu/resources/curriculum
- Slides

- http://www.slideshare.net/EUCLIDproject/querying-linked-data

- Webinar and screencasts
  - Webinar Part I https://vimeo.com/61618438
  - Webinar Part II https://vimeo.com/61618437
  - Screencast: Sesame (https://vimeo.com/61612180)
  - Screencast: Seevl https://vimeo.com/57931690
- Exercises
  - http://www.euclid-project.eu/node/49/take
- eBook chapter
  - HTML http://www.euclid-project.eu/modules/chapter2
  - iBook http://www.euclid-project.eu/sites/default/files/resources/Chapter2.ibooks
  - ePUB http://www.euclid-project.eu/sites/default/files/resources/Chapter2.epub
  - Kindle http://www.euclid-project.eu/sites/default/files/resources/Chapter2.azw3

2.3 Providing Linked Data

This module covers the whole spectrum of Linked Data production and exposure. After a grounding in the Linked Data principles and best practices, with special emphasis on the VoID vocabulary, we cover R2RML, operating on relational databases, Open Refine, operating on spreadsheets, and GATECloud, operating on natural language. Finally we describe the means to increase interlinkage between datasets, especially the use of tools like Silk.

We also paid special attention to providing examples with supporting tools and recorded a number of screencasts for this purpose.

2.3.1 Module Detailed Outline

All the materials to the third module are available at http://www.euclid-project.eu/#chapter3. The following listing describes the covered topics in detail.

3.1 Introduction and Motivation

3.2 Linked Data Lifecycle

- Linked Data Principles
- Tasks fro Providing Linked Data

3.3 Creating Linked Data

- Data extraction, giving names (URIs), selecting vocabularies

3.4 Interlinking Linked Data

- Link discovery
- Manual interlinking, automatic interlinking
- Interlinking with SKOS

3.5 Publishing Linked Data

- Describing dataset with metadata (VoID)
- Making the dataset accessible (dereferencing HTTP URIs, RDF dump, SPARQL endpoint, RDFa)
- Making the dataset searchable (search engine support)

- Exposing the dataset in repositories (creating new ones - CKAN, using the Data Hub, the Linking Open Data Cloud)

3.6 Linked Data Publishing Checklist

3.7 Tools for Providing Linked Data

- OpenRefine: Extracting data from spreadsheets
- R2RML: Extracting data from RDBMS
- GATECLOUD: Extracting data from text
- CALAIS: Extracting data from text
- Silk: Interlinking data sets

2.3.2 Available Materials

The currently available materials to module 3: Providing Linked Data can be found at:

- Outline
  - http://www.euclid-project.eu/resources/curriculum
- Slides
  - Module slides http://www.slideshare.net/EUCLIDproject/providing-linked-data
  - Presentation on Big Linked Data http://www.slideshare.net/EUCLIDproject/big-linked-data
  - Presentation on Creating Data Science Curriculum for Professionals http://www.slideshare.net/EUCLIDproject/data-science-curriculum-v-32
  - Presentation on Mapping Relational Databases to Linked Data http://www.slideshare.net/EUCLIDproject/r2-rml-londonsemweb201304
- Webinar and screencasts
  - Webinar Part I https://vimeo.com/64709409
  - Webinar Part II https://vimeo.com/64709408
  - Screencast: OpenRefine (https://vimeo.com/62430786)
- Exercises
  - http://www.euclid-project.eu/node/74/take
- eBook chapter
  - HTML http://www.euclid-project.eu/modules/chapter3
  - iBook http://www.euclid-project.eu/sites/default/files/resources/Chapter3.ibooks
  - ePUB http://www.euclid-project.eu/sites/default/files/resources/Chapter3.epub
  - Kindle http://www.euclid-project.eu/sites/default/files/resources/Chapter3.azw3

2.4 Interaction with Linked Data

This module focuses on providing means for exploring Linked Data. In particular, it gives an overview of current visualization tools and techniques, looking at semantic browsers and applications for presenting the data to the end user. We also describe existing search options, including faceted search, concept-based search and hybrid search, based on a mix of using semantic information and text processing. Finally, we conclude with approaches for

Linked Data analysis, describing how available data can be synthesized and processed in order to draw conclusions. The module includes a number of practical examples with available tools as well as an extensive demo based on analysing, visualizing and searching data from the music domain. The fourth module is very practice oriented and all of the though technologies are backed up by examples and tools.

2.4.1 Module Detailed Outline

The current set of available materials can be found at http://www.euclid-project.eu/#chapter4. The following listing describes the covered topics in detail.

4.1 Introduction and Motivations

4.2 Linked Data Visualisation

- Visualisation Techniques
    - Challenges for Linked Data Visualization
    - Classification of Visualization Techniques
    - Applications of Linked Data Visualization Techniques
- Linked Data Visualization Tools
    - Linked Data Visualization Tool Requirements
    - Linked Data Visualization Tool Types
    - Linked Data Visualization Examples: Sig.ma, Sindice, Information Workbench, LOD live, LOD Visualisation
- Linking Open Data Cloud Visualization
    - “The Linking Open Data cloud diagram” by Richard Cyganiak and Anja Jentzsch
    - “Linked Open Data Cloud” generated by Gephis
    - “Linked Open Data Graph” by Protovis
- LD Reporting
    - Google Webmaster Tool

4.3 Linked Data Search

- Semantic Search Process
- Semantic Search and Linked Data
    - Semantic Search vs. SPARQL query
    - Semantic Search with Google
    - Semantic Search with DuckDuckGo
- Faceted Search: Information Workbench, FacetedDBLP
- Classification of Search Engines
    - Semantic Data Search Engines: Swoogle, Watson
    - Searching for Vocabularies: LOV Portal
    - Searching for Documents: Semantic Web Search Engine (SWSE), Sindice

4.4 Methods for Linked Data Analysis

- Features of Linked Data analysis

- Data Aggregation and Filtering
- Statistical analysis: R for SPARQL
- Machine learning

2.4.2 Available Materials

The currently available materials to module 4: Interaction with Linked Data can be found at:

- Outline
  - http://www.euclid-project.eu/resources/curriculum
- Slides
  - Module slides http://www.slideshare.net/EUCLIDproject/interaction-with-linked-data
- Webinar and screencasts
  - Webinar Part I http://www.slideshare.net/EUCLIDproject/interaction-with-linked-data-part-i
  - Webinar Part II http://www.slideshare.net/EUCLIDproject/interaction-with-linked-data-part-ii
  - Screencast: Visualizing SPARQL Query Results (http://vimeo.com/68847721)
  - Screencast: Search Capabilities of the Information Workbench (http://vimeo.com/68847720)
  - Screencast: Interacting with Linked Data (http://vimeo.com/68847632)
- eBook chapter
  - HTML http://www.euclid-project.eu/modules/chapter4
  - iBook http://www.euclid-project.eu/sites/default/files/resources/Chapter4.ibooks
  - ePUB http://www.euclid-project.eu/sites/default/files/resources/Chapter4.epub
  - Kindle http://www.euclid-project.eu/sites/default/files/resources/Chapter4.azw3

2.5 Building Linked Data Applications

This module represents the first set of advanced topics, which are part of EUCLID's curriculum. The fifth module gives details on technologies and approaches towards exploiting Linked Data by building Linked Data applications. In particular, it gives an overview of popular existing applications and introduces the main technologies that support implementation and development. Furthermore, it illustrates how data exposed through common Web APIs can be integrated with Linked Data in order to create mashups.

Similarly to the previous module, this module is very practice oriented and we are planning on building on the examples and demo system developed for Interacting with Linked Data, in order to demonstrate how applications can be designed and implemented on top of the available data.

2.5.1 Module Detailed Outline

The current set of available materials can be found at http://www.euclid-project.eu/#chapter5. The following listing describes the covered topics in detail.

5.1 Introduction and Motivations

5.2 Linked Data Applications

- Characterization of Linked Data Applications
- Categories of Linked Data Applications

- Examples: Data.gov.uk, Data.gov, BBC – Dynamic Semantic Publishing, ResearchSpace, Open Pharmacology Space, Information Workbench, eCloudManager – Integrated View on the Data Center

5.3 Using Web APIs

- Underlying Technology Basics
- Web APIs - Motivation
- Richardson Maturity Model for REST Services
- Well-Known Web APIs: Freebase API, Twitter API, Last.fm API, Foursquare API, Amazon S3

5.4 Linked Data application architecture

- Software Architecture
    - Client-Server Model
    - Multitier Architecture
- Architecture of Linked Data Applications
    - Linked Data Architectural Patterns
    - General Architecture of Linked Data Applications
        - Publication Layer
        - Data Access Component
        - Data Integration Component
        - Data Layer
        - Application and Presentation Layers
- Challenges for Developing Linked Data Applications

5.5 Linked Data application development frameworks

- Information Workbench Architecture
    - Ontology as a "structural backbone" of the application
    - Managing data
        - Connecting to the data repository
        - Integrating external data using data providers
        - Federated data access
    - Creating the user interface
        - Data-driven UI: Providing views over data resources
        - Ontology-driven UI structure:
            - Class views: providing overview over the dataset
            - Wiki templates: common UI structure for data instances
    - Constructing a mashup
    - Data authoring

2.5.2 Available Materials

The currently available materials to module 5: Building Linked Data Applications can be found at:

- Outline

- http://www.euclid-project.eu/resources/curriculum

- Slides
  - Module slides http://www.slideshare.net/EUCLIDproject/building-linked-data-applications-27768679

- Webinar and screencasts
  - Webinar Part I http://www.slideshare.net/EUCLIDproject/building-linked-data-applications-part-i
  - Webinar Part II http://www.slideshare.net/EUCLIDproject/building-linked-data-applications-part-ii

Currently we are in the process of preparing the screencasts and the first version of the written chapter. According to our time plan, the work on this module will be completed by mid December.

2.6 Scaling-up

The final module of the curriculum is devoted to dealing with large amounts of semantic data and using and managing these in an effective way. In particular, this module is centred around scaling the reasoning approaches for RDF, scaling the data storage solutions and discussing the current ways for accessing and generating big volumes of semantic data. In summary, it addresses the main issues of Linked Data and scalability. It is important to point out that this module reflects on the latest developments in the filed. Therefore, a lot of the listed work is based on pioneer approaches, which might not have achieved a wider adoption jet. This module is the one that is the least based on directly used tools and commonly plied practices. However, the discussed topics have only recently been actively addressed and are important in the overall context of gaining skills and expertise for dealing with Linked Data.

2.6.1 Module Detailed Outline

The module six on 'Scaling-up' covers the topics given in the following listing. All the available materials will be provided at http://www.euclid-project.eu/#chapter6. Currently we have just started with collecting the related materials and finalising the detailed outline of the module.

6.1 Introduction

- Challenges of dealing with big volumes of Linked Data

6.2 Scaling Reasoning for Linked Data

- RDFS reasoning with Hadoop
- RDFS reasoning with GPUs

6.3 Scaling Storage for Linked Data

- NoSQL databases for RDF management
- Storing big volumes of RDF data

6.4 Generating and Accessing Big Volumes of Semantic Data

- Streaming SPARQL
- Semantic sensors data

Currently, the covered topics and content of all modules are fixed. Naturally, the final modules are not as rich in terms of materials yet, but as we progress, all modules will be covered with a list on available training resources. We plan to complete the production of all main materials by the end of early 2014. This would allow for enough time for sharing and announcing the achieved results.

In particular, we will complete the work on the final module in accordance with the following time plan:

- 15.11 first version of slides

- end of November internal webinar, feedback
- 02.12 second version of slides
- 09.12 review of slides and chapter done
- 09.12 exercises and examples finalized
- 16.12 final version of slides
- mid-end of December public webinar (19.12.2013)
- 15.01 first version of chapter, Maria Esther starts reading
- 30.01 review of chapter
- 30.01 Screencasts and multiple choice questions
- 21.02 (approx.) final version of slides, final version of HTML published on website

3 Alignment with Related Training Activities

With the growing use and application of the Linked Data principles and related technologies, it is only natural that there are already some teaching materials and courses on the topic. In order to be able to provide a curriculum that is up to date but that is also competitive in terms of the offered training, we identify related training activities and describe how they align with the courses offered by the project. Based on this analysis we can objectively argue for the completeness of the provided curriculum, in terms of the covered topics, but also taking into consideration the training goals and target audience. Furthermore, we can better align with existing initiatives in the flied, initiate collaborations and plan the organisation of the training events accordingly.

In addition, this section we analyse and compare existing training activities and curricula in the area of Linked Data and Data Science in general. Deliverable 1.1.1 provided an initial overview of some of the related activities, here we extend these and add some further ones. In addition, we provide a set of learning pathways for acquiring skills in four main expertise directions – data architect, data manager, data analyst and data application developer. We demonstrate how different modules from different curricula can be combined in order to gain a cross-domain knowledge, which is commonly required by data practitioners.

3.1 Areas of Expertise for Data Professionals

We define four main areas of expertise, in which skills can be acquired, either by using only the EUCLID training modules or by combining them with materials from further existing training curricula in the fields of Linked Data and Data Science in general. In particular we consider the following areas of expertise:

- **Data Architect:**

A data architect is someone who is concerned with designing the structure of the data, determining, the storage and deployments solutions, the technological principles, and the used data model. The data architect has also some expertise in the area of converting the data in the required structure, in the context of Linked Data this would include publishing Linked Data, determining which links are going to be created and which properties need to be linked. The data architect also has expertise in the field of converting existing data into Linked Data, in the context of publishing Linked Data.

- **Data Manager:**

The data manager is concerned with gardening the data, performing and committing updates, optimising the storage solution. The main goal of the data manager is to keep the data up to date and to make sure that it is stored, distributed and accessible in a scalable and flexible way. Naturally, this includes some querying, browsing and searching skills in order to be able to get a good overview of the current state.

- **Data Analyst:**

The data analyst performs search, browsing, more complex queries and uses visualisation tools in order to explore the data, makes statements about important topics, trends and correlations. In this context, it is important to be able to draw conclusions, based on the available datasets, identify tends and extract fact that serves as input to the decision making process. Supporting tools play an important role in this context.

- **Data Application Developer**

The data application developer builds applications on top of the data, by creating user interfaces for exploring it, combining it with Web APIs as further data sources, or integrating it as part of processing components. He/she needs to have comprehensive and overall competencies, as well as more thorough knowledge about how to proces and transform the data.

The following table (see Table 1) describes, which Euclid training modules are most relevant for which data professional training. Naturally, the introductory and intermediate modules are suitable for all data expert types, even if certain topics might be more suitable for a given audience than others.

	Data Architect	**Data Manager**	**Data Analyst**	**Data Application Developer**
Introductory Level	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios

Intermediate Level		Module 2: Querying Linked Data	Module 2: Querying Linked Data	Module 2: Querying Linked Data
	Module 3: Providing Linked Data	Module 3: Providing Linked Data		Module 3: Providing Linked Data
			Module 4: Interaction with Linked Data	Module 4: Interaction with Linked Data
Advanced Level				Module 5: Creating Linked Data Applications
	Module 6: Scaling up	Module 6: Scaling up		Module 6: Scaling up

Table 1: EUCLID Modules for Skills Development

3.2 PlanetData Training Curriculum

In this section we shortly revisit the PlanetData curriculum, the offered topics and available materials. The PlanetData project aims to establish a community of researchers that supports organizations in exposing their data in new and useful ways. In this context, the curriculum covers four main topics – Semantic Technology, Database Technology, Linked Data and Data Streams. The produced training materials are available as part of a series of video lectures http://videolectures.net/planetdata_training_curriculum/

From the topics covered by the PlanetData curriculum, the Linked Data topic is the most relevant one for EUCLID's curriculum. However, some of the remaining topics can be used to determine partial coverage of such as Semantic Technologies and module 1, which introduces all the main underlying principles. The complete Planet Data curriculum is available at[7], while the covered Linked Data topics are listed below.

1. Introduction
    a. The Data Deluge
    b. The Rationale for Linked Data
    c. Intended Audience
2. Principles of Linked Data
    a. The Principles in a Nutshell
    b. Naming Things with URIs
    c. Making URIs Defererencable
    d. Providing Useful RDF Information
    e. Including Links to other Things
3. The Web of Data
    a. Bootstrapping the Web of Data
    b. Topology of the Web of Data
4. Linked Data Design Considerations
    a. Using URIs as Names for Things
    b. Describing Things with RDF

[7] http://www.planet-data.eu/sites/default/files/pr-material/deliverables/D6.1_Training_curriculum.pdf

c. Publishing Data about Data
d. Choosing and Using Vocabularies
e. Making Links with RDF

5. Recipes for Publishing Linked Data
    a. Linked Data Publishing Patterns
    b. The Recipes
    c. Additional Approaches to Publishing Linked Data
    d. Testing and Debugging Linked Data
    e. Linked Data Publishing Checklist
6. Consuming Linked Data
    a. Deployed Linked Data Applications
    b. Architecture of Linked Data Applications
    c. Effort Distribution between Publishers, Consumers and Third

For the module on providing Linked Data, some on the topics from section 4. Recipes for Publishing Linked Data were taken into consideration. Section 1 trough 3 were partially covered by the introductory module, where we describe the main relevant technologies, the underlying principles and the vision of the Web of Data. Finally, section 5 on consuming Linked Data is somewhat similar to module 5, which describes how to build applications on top of Linked Data. The main difference here is that we take a very practical approach by giving specific examples, set of tools and solutions that can directly be used. We do not focus so much on the architectural approach or the individual roles of the parties taking part in the data consumption and production.

3.2.1 Skills Development Based on the PlanetData Curriculum

In this section we describe how the PlanetData training sections can be combined with the individual modules, in order to achieve competencies in the four main fields of expertise – data architect, data manager, data analyst and data applications developer.

In direct comparison, the EUCLID curriculum represents a more practice-oriented approach towards covering some of the topics, which are also part of the PlanetData curriculum. Each of our modules directly points out tools, technologies and approaches and goes in depth, when it comes to addressing specific challenges that a data practitioner might face. It seems that the PlanetData curriculum focuses on providing more insight and detail on the introductory topics and takes the more advanced sections only up to a certain theoretical level. Therefore, we can conclude that it is foreseen for a rather broader and inexperienced audience, in contrast to the EUCLID's curriculum, which addresses firstly the needs of data professionals, while still be in suitable for a more general training.

Table 2 visualises the suggested way of combining the PlanetData sections in order to gain expertise in a certain field of data expertise. As it can be seen by comparing Table 1 and Table 2, the content is not really complimentary but there is some overlap in the topics. Therefore, this curriculum should not really be used in conjunction with EUCLID materials but rather as a way of crosschecking that all relevant topics have been covered.

	Data Architect	**Data Manager**	**Data Analyst**	**Data Application Developer**
Introductory Level	1. Principles of Linked Data	1. Principles of Linked Data	1. Principles of Linked Data	1. Principles of Linked Data
	2. The Web of Data	2. The Web of Data	2. The Web of Data	2. The Web of Data
Intermediate	3. Linked Data Design	3. Linked Data Design		

Level	Considerations	Considerations		
	4. Recipes for Publishing Linked Data	4. Recipes for Publishing Linked Data		
Advanced Level				5. Consuming Linked Data

Table 2: Skills Alignment with the PlanetData Curriculum

3.3 Open Data Institute (ODI)

One of the important developments that have taken place in the context of teaching and Linked Data is the launching of activities of the Open Data Institute (ODI)[8]. The ODI is an independent, non-profit, non-partisan, limited by guarantee company, which aims to "catalyse an open data culture that has economic, environmental and social benefits". In the context of this goal, the ODI offers a number of training events such as the Open Data in Practice[9], which is a five-day course and the Introduction to Open Data for Journalists: Finding Stories in Data[10].

The Open Data in Practice course is of particular relevance to EUCLID's curriculum, since it provides a very business-oriented and practical introduction to open data. In particular, it gives an introduction to the technical, commercial and legal aspects of open data. Furthermore, it aims to support the uptake of the principles by highlight key opportunities for working with open data and how they can be exploited across government, business and society. The course is designed to enable anyone to understand how to publish, consume and exploit open data as well as give an understanding of best practice, law, licensing and policy issues.

The description of the course states that it is targeted towards open data practitioners, policy officials and advisors, account and project managers, statisticians and analysts, strategists, entrepreneurs, business developers, ICT suppliers, knowledge managers, policy owners, developers, information architects, journalists, research and intelligence. Similarly to the EUCLID curriculum, no previous experience is required but good computer skills and familiarity with information technologies is of benefit.

The outline of the Open Data in Practice course is the following:

1. What is Open Data? What are the benefits? Considering personal data?
2. Let's make some data. How does it work on the web?
3. Licensing, the law and best practice
4. From publishing to consuming open data, including open data standards
5. Tools for analysing data: cleaning, validating, and enriching data.
6. Analysing data continued: Establishing trust Visualising data
7. Business and open data: benefits, applications, value propositions
8. Innovating with open data
9. Hack Day in teams

A it can be seen the course is very practice oriented, talking specific topics such as licensing, law and best practise. Furthermore, it also includes implementation sessions every day, so that the participants can directly apply the theoretical knowledge. Section 2 is similar to module 3 on providing Linked Data. Furthermore, sections 5 and 6 relate to our module on 'Interaction with Linked Data', including analysing and visualising data. What we have not reflected on in EUCLID's curriculum is the framework around dealing with data, which includes legal issues and

[8] http://www.theodi.org
[9] http://www.theodi.org/courses/open-data-practice
[10] http://www.journalism.co.uk/introduction-course-open-data-for-journalists/s382/

business models that can be applied. Therefore, it seems that the ODI addresses not simply data practitioners but also managers and decision makers. This is important when designing a course that tries not only to provide an introduction to the technology but also to encourage its adoption. In contrast, our curriculum is focused on developing practical skills and not so much on motivating the need for Linked Data and demonstrating how it can be economically exploited.

Table 3 visualises a possible combination of the EUCLID's modules with the ODI curriculum, in the predefined four areas of expertise. It is important to point that ODI's training is also targeted towards project managers and decision makes, especially in the context of reflecting on licensing issues, laws and best practices. Therefore, it has a business-setting oriented content, as opposed to addressing solely data practitioners.

	Data Architect	**Data Manager**	**Data Analyst**	**Data Application Developer**
Introductory Level	Module 1: Introduction and Application Scenarios OR	Module 1: Introduction and Application Scenarios OR	Module 1: Introduction and Application Scenarios OR	Module 1: Introduction and Application Scenarios OR
	ODI 1. What is Open Data? What are the benefits? Considering personal data?	ODI 1. What is Open Data? What are the benefits? Considering personal data?	ODI 1. What is Open Data? What are the benefits? Considering personal data?	ODI 1. What is Open Data? What are the benefits? Considering personal data?
	ODI 2. Let's make some data. How does it work on the web?	ODI 2. Let's make some data. How does it work on the web?	ODI 2. Let's make some data. How does it work on the web?	ODI 2. Let's make some data. How does it work on the web?
Intermediate Level		Module 2: Querying Linked Data	Module 2: Querying Linked Data	Module 2: Querying Linked Data
	Module 3: Providing Linked Data	Module 3: Providing Linked Data		Module 3: Providing Linked Data
	ODI 4. From publishing to consuming open data, including open data standards	ODI 4. From publishing to consuming open data, including open data standards		
	ODI 3. Licensing, the law and best practice	ODI 3. Licensing, the law and best practice		ODI 3. Licensing, the law and best practice
		ODI 5. Tools for analysing data: cleaning, validating, and enriching data	ODI 5. Tools for analysing data: cleaning, validating, and enriching data	
			Module 4: Interaction with Linked Data	Module 4: Interaction with Linked Data
			ODI 6. Analysing data continued: Visualising data	ODI 6. Analysing data continued: Visualising data
	ODI 7. Business and open data: benefits, applications, value propositions			

	ODI 8. Innovating with open data			
Advanced Level				Module 5: Creating Linked Data Applications
	Module 6: Scaling up	Module 6: Scaling up		Module 6: Scaling up

Table 3: Skills Alignment with ODI's Curriculum

As stated by the course description the objectives and expected competencies of the ODI curriculum are the following:

- Have an overview of open data, law, web technologies and its application potential
- Have an understanding of the architecture and openness of the web
- Understand the considerations of publishing personal data
- Have a practical understanding of how to publish, consume, and exploit open data
- Understand processes required to release large data sets
- Have developed the ability to evaluate open data strategies
- Have an increased knowledge of how to commercialise and innovate using open data
- Have worked with others to produce data
- Understand the vocabulary around data, such as linked data and the semantic web
- Have developed an ability to share and brief others on benefits of open data

These objectives once again emphasise the practical orientation of the course and also confirm that further topics such as the legal issues or business setting should be included in curricula that support training that supports the adoption of a particular technology.

3.4 Lean Semantic Web

Jie Bao[11] offers a tutorial on Lean Semantic Web[12] including a list of the covered topics. The curriculum is extensive and in addition to covering the semantic technologies, also provides a lot of details on data management, data storage and databases. Overall the sections are very data-centric and focused on explaining in detail the related technologies. A list of all the sections is provided below.

Section 1 Data Representation

- What is data and structure?
- Value of unstructured data
- Value of structured data
- Cost of data modeling
- The art of readable knowledge
- Naming resources
- Locating resources
- Relating resources
- Syntax: XML, JSON, YAML, RDF, Python etc.
- Implementation: some Python (RDFLib...)

Section 2 Databases

- Cost of modeling and indexing semantics in database
- Use relational DB for semantic modeling
- Document database: MongoDB, Elastic Search etc

[11] http://www.baojie.org/
[12] https://github.com/baojie/leansemanticweb/blob/master/Syllabus.md

- Graph database: TinkerPop stack, Neo4j, OrientDB
- Graph batch processing: Pregel, Hama, GraphChi etc
- Querying triples (with RDB, document db, graph db, or dedicated triple store)
- Implementation: some Python

Section 3 Search and Findability

- Database vs search engine
- Cost of inverted index
- Extend inverted index to model semantic relations
- Understanding user queries: from keywords to sentences
- Faceted search: Elastic Search and Solr
- Graph search
- Implementation: some Python (ESClient...)

Section 4 Data Exchange and Integration

- Portablility
- Protocol Buffers and Thrift
- Email and MIME
- JSON-RPC
- XMPP and Google Wave Protocol
- REST API design
- Some most important data APIs
- Implementation: some Python (JSON...)

Section 5 Inference

- Cost of reasoner and index (IR, DB, KB)
- Just-in-time knowledge
- Practical rule modeling
- Inference as graph operations
- Inference using databases
- Inference using functional programming
- Implementation: some Python (Pydatalog, Fuxi)

Section 6 Knowledge Extraction

- Cost of knowledge extraction
- Data cleaning
- Structure extraction
- Shallow parsing
- Entity extraction
- Relation extraction
- Implementation: some Python (NLTK...)

Section 7 Visualization

- Cognitive background
- Exhibit and others
- D3 (and other JavaScript lib)
- NetworkX and (and other Python lib)
- Implementation: some Python (mapplotlib...) and Javascript

Section 8 User Interaction

- It's about people, not machine
- Guided data exploration and discovery: why semantics is part of the solution
- Query formulating
- Faceted Browser
- Mobile search, Voice interface and personal assistants
- Implementation: some Python (ElasticSearch, Redis, RDF store based)

Section 9 Big Data and Lean Data

- Measuring semantics in data
- Small is beautiful in knowledge
- The rule of knowledge growth on Web
- Small knowledge: in-memory graph models
- Big knowledge: knowledge bases on clusters
- Big data
- Datasets: Freebase, DBPedia, LOGD, Factual etc.
- Platforms: EC2, and some others
- Implementation: some Python (boto, starcluster)

Section 10 Lean Application Development

- Build, measure, learn
- Lean Canvas
- MVP
- Build: mockup strategies, pretotyping, prototyping
- Measure: key metrics, but not vanity metrics
- Learn: why^5
- Semantic Wordpress/Drupal/Wiki, etc,

The list of covered topics is long and there are a lot of details related to the specific underlying technologies. Overall, the curriculum is targeted towards developers who have to deal with different challenges related to data. The sections are not necessarily based on the different data-related tasks that have to be completed rather on the development that can be done for and with data (such as databases or applications).

This curriculum provided us with some interesting insights about topics or motivation that needed to be included in EUCLID's curriculum as well. Section 1 is similar to module 1, giving an introduction to the basic underlying principles. The main difference is that we did not motivate the need for structured data and the resulting benefits in so much detail. This has again to do with the target audience of EUCLID's curriculum and with the fact that is it very practically-oriented, thus assuming that the participants are already interested in Linked Data and do not have to be persuaded first. Section 2 can be seen as a prerequisite to our curriculum, as there is no module that focuses explicitly on databases. Instead databases are mentioned as tools towards achieving certain tasks such querying, publishing, analyzing, visualizing, etc.

Section 3 was actually taken into account while determining the content of modules 2 and 4 on 'Querying Linked Data' and on 'Interacting with Linked Data'. Indexing plays an important part for supporting search and we included examples of different systems (e.g. Solr). Still the overall level of the section is higher than we aim to cover while providing practical training on how to query and search Linked Data. The technologies mentioned in Section 4 are used as a basis for explaining the development of applications on top of Linked Data, which will be covered in EUCLID in module 5. Yet again, the level of detail here is not reflected in the module, since we only aim to describe the fundamentals and focus on the development scenarios.

Section 5 is included in module 2 as part of more complex ways of querying Linked Data, while Section 6 is partially reflected as part of the steps for providing Linked Data. Similarly to previous sections, our curriculum focuses on describing how Linked Data can be created and while knowledge extraction topics are relevant, we do not reflect on them in so much detail.

Sections 7 and 8 are covered as part of the module on interaction with Linked Data. EUCLID's curriculum covers more tools and approaches than the here mentioned ones. In fact, we are aiming to provide a catalog of tools for manipulating Linked Data, including visualization. The module is very important, since it gives the basics for exploring the data and benefiting from it. Therefore this is one of the few cases where EUCLID will provide mode details than the here discussed sections.

We envision to cover the topics listed in Sections 9 under module 6 on scaling-up. In fact this section was quire useful, since we were able to update our initial outline. Finally Section 10 will be integrated as part of module 5 on building applications on top of Linked Data.

The following table (see Table 4) visualizes the alignment of the Lean Semantic Web curriculum with EUCLID's curriculum. As it can be seen, there is a slight overlap, however, the two curricula are quire complimentary, especially in the context of adding more data fundamentals such as data structures and databases, and in the context of providing more general data knowledge, as opposed to focusing strictly on Linked Data. Furthermore, by combining learning blocks from the two training plans, the two tracks for data architect and data analyst are covered well. The only topics that are still missing, in order to make complete curricula, are data extraction and natural language processing (data architect), and data processing and mining (data analyst).

The target audience of the curriculum presented in this section is technology experts in the area of data storage and use, with focus on tools, technologies and practical skills. In comparison to EUCLID's curriculum, the scope is more data, technology and implementation-centric, not so much semantics, fundamentals and supporting the completion of common tasks, which have to be performed by data practitioners.

	Data Architect	**Data Manager**	**Data Analyst**	**Data Application Developer**
Introductory Level	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios
	LSW Section 1: Data Representation	LSW Section 1: Data Representation	LSW Section 1: Data Representation	LSW Section 1: Data Representation
	LSW Section 2: Databases	LSW Section 2: Databases	LSW Section 2: Databases	LSW Section 2: Databases
Intermediate Level		Module 2: Querying Linked Data	Module 2: Querying Linked Data	Module 2: Querying Linked Data
	LSW Section 3: Search and Findability	LSW Section 3: Search and Findability		LSW Section 3: Search and Findability
		LSW Section 5: Inference	LSW Section 5: Inference	LSW Section 5: Inference
	Module 3: Providing Linked Data	Module 3: Providing Linked Data		Module 3: Providing Linked Data
	LSW Section 6: Knowledge Extraction	LSW Section 6: Knowledge Extraction		LSW Section 6: Knowledge Extraction
			Module 4: Interaction with Linked Data	Module 4: Interaction with Linked Data
			LSW Section 7: Visualization	LSW Section 7: Visualization
			LSW Section 8: User Interaction	LSW Section 8: User Interaction
Advanced Level				Module 5: Creating Linked Data Applications
				LSW Section 4: Data Exchange and Integration
				LSW Section 10: Lean Application Development

	Module 6: Scaling up	Module 6: Scaling up		Module 6: Scaling up
	LSW Section 9: Big Data and Lean Data	LSW Section 9: Big Data and Lean Data		LSW Section 9: Big Data and Lean Data

Table 4: Skills Alignment with Lean Semantic Web Curriculum

Overall, this curriculum was very helpful in making sure that we have covered a sufficient set of topics, in the required level of detail. It also became evident that it might be difficult to find the balance between developing a curriculum based on theoretical approaches and one focused on implementation and specific technologies. By focusing on the tasks that need to be completed by data practitioners, we always aim to provide the required knowledge and skills in the corresponding modules, while not relying too much on specific technologies which are often subject to frequent changes.

3.5 Cloudera Data Scientist Curriculum

Cloudera[13] develops open-source software for a Big Data, with its most prominent product being Hadoop – a data storage and management solution. Cloudera aims to support companies in interacting with large datasets at high speed, providing management and analysis functionalities on top. As part of its recent initiatives Cloudera has started an online university program for training. One of the main training courses offered is the program in data science[14]. The aim is to develop practical skills, which can be applied in real-world conditions, for designing and developing a production-ready data science solution. The course is still in the beta version and is undergoing a process of verification and refinement. Currently, candidates must pass a written exam in data science essentials and complete a data science problem based on a real-world scenario.

The course offers a curriculum based on 11 sections and a rich collection of related materials, including books, papers, courses, presentations, blogs, meetups, etc. Below we give more details on the covered topics in the Data Science Essentials Study Guide[15]:

1. Data Acquisition
    a. Access and load data from a variety of sources into a Hadoop cluster, including from databases and systems such as OLTP and OLAP as well as log files and documents.
    b. Deploy a variety of acquisition techniques for acquiring data, including database integration, working with APIs
    c. Use command line tools such wget and curl; Use Hadoop tools such as Sqoop and Flume
2. Data Evaluation
    a. Knowledge of the file types commonly used for input and output and the advantages and disadvantages of each
    b. Methods for working with various file formats including binary files, JSON, XML, and .csv
    c. Tools, techniques, and utilities for evaluating data from the command line and at scale
    d. An understanding of sampling and filtering techniques
    e. A familiarity with Hadoop SequenceFiles and serialization using Avro
3. Data Transformation
    a. Write a map-only Hadoop Streaming job
    b. Write a script that receives records on stdin and write them to stdout

[13] http://www.cloudera.com
[14] http://university.cloudera.com/certification/ccp.html
[15] http://university.cloudera.com/certification/prep/datascience.html

   c. Invoke Unix tools to convert file formats
   d. Join data sets; Write scripts to anonymize data sets
   e. Write a Mapper using Python and invoke via Hadoop streaming
   f. Write a custom subclass of FileOutputFormat; Write records into a new format such AvroOutputFormat or SequenceFileOutputFormat

4. Machine Learning Basics
   a. Understand how to use Mappers and Reducers to create predictive models
   b. Understand the different kinds of machine learning, including supervised and unsupervised learning
   c. Recognize appropriate uses of the following: parametric/non-parametric algorithms, support vector machines, kernels, neural networks, clustering, dimensionality reduction, and recommender systems
5. Clustering
   a. Define clustering and identify appropriate use cases
   b. Identify appropriate uses of various models including centroid, distribution, density, group, and graph
   c. Describe the value and use of similarity metrics including Pearson correlation, Euclidean distance, and block distance
   d. Identify the algorithms applicable to each model (k-means, SVD/PCA, etc.)
6. Classification
   a. Describe the steps for training a set of data in order to identify new data based on known data
   b. Identify the use cases for logistic regression, Bayes theorem
   c. Define classification techniques and formulas
7. Collaborative Filtering
   a. Identify the use of user-based and item-based collaborative filtering techniques
   b. Describe the limitations and strengths of collaborative filtering techniques
   c. Given a scenario, determine the appropriate collaborative filtering implementation
   d. Given a scenario, determine the metrics one should use to evaluate the accuracy of a recommender system
8. Model/Feature Selection
   a. Describe the role and function of feature selection
   b. Analyze a scenario and determine the appropriate features and attributes to select
   c. Analyze a scenario and determine the methods to deploy for optimal feature selection
9. Probability
   a. Analyze a scenario and determine the likelihood of a particular outcome
   b. Determine sample percentiles; Determine a range of items based on a sample probability density function
   c. Summarize a distribution of sample numbers
10. Visualization
   a. Determine the most effective visualization for a given problem

b. Analyze a data visualization and interpret its meaning

11. Optimization
    a. Understand optimization methods
    b. Identify 1st order and 2nd order optimization techniques
    c. Determine the learning rate for a particular algorithm
    d. Determine the sources of errors in a model

As it can be seen, the offered training is extensive, covering a broader range of topics, starting with data fundamentals, including machine learning approaches, and concluding with optimisation. Furthermore, the topics are presented in a very practice-oriented way, directly relating to the used tools. The only drawback is that the course is, understandably, bound to the solutions and products offered by Cloudera, focusing on Hadoop. Still, the topics complement the modules offered by EUCLID's curriculum well, providing a line of technology-specific learning paths in the context of Big Data.

Table 5 visualises how our training modules can be combined with sections of the Cloudera data scientist training course. It is important to point out that none of the topics are really at introductory level (for example, introduction to big data or introduction to data formats), starting directly with providing some details on advanced skills. Furthermore, similarly to our approach, the topics are defined as tasks that need to be completed by a trained data practitioner – filtering, classification, clustering, visualisation. This training approach ensures a very practical skills development process.

	Data Architect	**Data Manager**	**Data Analyst**	**Data Application Developer**
Introductory Level	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios
Intermediate Level		Module 2: Querying Linked Data	Module 2: Querying Linked Data	Module 2: Querying Linked Data
	Module 3: Providing Linked Data	Module 3: Providing Linked Data		Module 3: Providing Linked Data
	Cloudera Section 1. Data Acquisition	Cloudera Section 1. Data Acquisition		
	Cloudera Section 2. Data Evaluation	Cloudera Section 2. Data Evaluation	Cloudera Section 2. Data Evaluation	
		Cloudera Section 3. Data Transformation	Cloudera Section 3. Data Transformation	
			Module 4: Interaction with Linked Data	Module 4: Interaction with Linked Data
			Cloudera Section 10. Visualisation	Cloudera Section 10. Visualisation
Advanced Level			Cloudera Section 4. Machine Learning Basics	
			Cloudera Section 5. Clustering	
			Cloudera Section 6.	

			Classification	
			Cloudera Section 7. Collaborative Filtering	
			Cloudera Section 8. Model/Feature Selection	
			Cloudera Section 9. Probability	
			Cloudera Section 11. Optimization	
				Module 5: Creating Linked Data Applications
	Module 6: Scaling up	Module 6: Scaling up		Module 6: Scaling up

Table 5: Skills Alignment with the Cloudera Data Scientist Curriculum

As it can be seen, the Cloudera training can be combined with EUCLID's modules in order to train data analysts, who can load and transform the data, apply different processing techniques such as filtering or classification, and even improve the used computational algorithms (optimisation). If optimisation also covers topics related to improving the data quality, structure or querying speed, this topic would also be relevant for data managers (this is why it is put in brackets). Overall, the Cloudera course aims to train specialists who focus on managing and processing the stored data. Therefore, the Cloudera and EUCLID curricula complement each other well.

3.6 EMC Data Science and Big Data Analytics Curriculum

ECM[16] is an IT provider for data storage solutions, including services for backup and recovery, cloud, big data, archiving, content managements, infrastructure management and security. The company has an educational portal, providing a course on data science and big data analytics[17]. The course aims to address the need for trained professionals, who are able to effectively use large amounts of data. This includes both skills as well as tool support. The training is focused on concepts and principles applicable to any technology environment and industry.

The Data Science and Big Data Analytics consist of six modules, focusing on practice-oriented data analytics. The covered topics are listed below:

1. Introduction to Big Data Analytics
    a. Big Data Overview
    b. State of practice in analysis
    c. The role of the Data Scientist
    d. Big Data Analytics in the industry verticals
2. End-to-end data analytics lifecycle
    a. Key roles for the successful analytics project
    b. Main phases of the lifecycle

[16] http://www.emc.com
[17] https://education.emc.com/content/_common/docs/certification/DSBDA_datasheet.pdf

   c. Developing core deliverables for stakeholders
3. Using R to execute basic analytics methods
   a. Intro to R
   b. Analysing and exploring data with R
   c. Statistics for model building and evaluation
4. Advanced analytics and statistical modeling for Big Data – Theory and Methods
   a. K-Means Clustering
   b. Association Rules
   c. Linear and Logistic Regression
   d. Naïve Bayesian Classifier
   e. Decision Trees
   f. Time Series analysis
   g. Text Analysis
5. Advanced analytics and statistical modeling for Big Data – Technology and Tools
   a. Using MapReduce/Hadoop for analysing unstructured data
   b. Hadoop ecosystem of tools
   c. In-database Analytics
   d. MADlib and Advanced SQL Techniques
6. Putting it all together
   a. How to operationalize an analytics project
   b. Creating the Final Deliverables
   c. Data Visualisation Techniques
   d. Hands-on Application of analytics Lifecycle to a Big Data Analytics Problem

The training is designed for business and data analysts, who are aiming to gain big data analytics skills. It is also relevant for managers, who use data analysis for decision-making. Overall, the course is relevant for anyone who is looking to enrich his/her analytic skills.

	Data Architect	**Data Manager**	**Data Analyst**	**Data Application Developer**
Introductory Level	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios
		EMC Section 1. Introduction to Big Data Analytics	EMC Section 1. Introduction to Big Data Analytics	
Intermediate Level		Module 2: Querying Linked Data	Module 2: Querying Linked Data	Module 2: Querying Linked Data
	Module 3: Providing Linked	Module 3: Providing Linked		Module 3: Providing

	Data	Data		Linked Data
			Module 4: Interaction with Linked Data	Module 4: Interaction with Linked Data
		EMC Section 2. End-to-end data analytics lifecycle	EMC Section 2. End-to-end data analytics lifecycle	
			EMC Section 3. Using R to execute basic analytics methods	
Advanced Level				Module 5: Creating Linked Data Applications
			EMC Section 4. Advanced analytics and statistical modeling for Big Data – Theory and Methods	
			EMC Section 5. Advanced analytics and statistical modeling for Big Data – Technology and Tools	EMC Section 5. Advanced analytics and statistical modeling for Big Data – Technology and Tools
			EMC Section 6. Putting it all together	EMC Section 6. Putting it all together
	Module 6: Scaling up	Module 6: Scaling up		Module 6: Scaling up

Table 6: Skills Alignment with the EMC Data Science and Big Data Analytics Curriculum

In the table above (see Table 6) we provide an overview how the EMC training can be combined with EUCLID modules in order to train data professionals. Similarly, to the courses offered by Cloudera, here we see that the learning path of a data analysis is very well covered. This is naturally to be expected, since this is also the main focus of the course. However, it is important to point out that by adding modules on querying Linked Data and interacting with Linked Data, the trainee will get a broader understanding of the current state of recent developments in the filed of data and profit from the connection of Linked Data and Big Data technologies and principles. Therefore, the two curricula are quire complementary.

3.7 GATE Training

GATE is an extensive open software framework, which supports a multitude of text-processing tasks. It covers a number of different languages, and, among others, enables text mining, web mining, information extraction, and semantic annotation. A series of different analysis and computation activities can be combined by defining a text processing workflow, based on integrating already preimplemented solutions or plugging-in own implementations. GATE is already widely used by research labs and universities, as well as by corporations and SMEs, worldwide. As part of providing support and training for the developed tools, GATE has offers courses, which are based on a curriculum consisting of three tracks. The tracks cover introductory level, basic programming skills, and advanced level. In addition, the curriculum includes two add-on courses, one of which is based directly on EUCLID

materials[18]. The reuse of some of the GATE training module can be very helpful in the context of developing skills within the four areas of knowledge expertise, which we have defined, and especially for data analysts.

In the following we list the GATE training modules[19]:

Track 1: Introduction to GATE and Text Mining

- Module 1: Introduction to GATE Developer
    - GATE concepts, finding your way around the GUI, loading and using existing processing resources and plugins
    - Loading, annotating and viewing existing language resources, creating and using applications
- Module 2: Information Extraction and ANNIE
    - Basic introduction to Information Extraction, running and evaluating an information extraction project
    - Using and customising ANNIE, GATE's IE tool
    - Using the Corpus QA and other evaluation tools
    - Introduction to semantic annotation with ontologies
- Module 3: Introduction to JAPE
    - Using JAPE grammars for annotation manipulation, using JAPE for named entity recognition
- Module 4: Introduction to Teamware, GateCloud, and Mímir
    - Teamware: A web-based collaborative corpus annotation tool
    - GateCloud: Low-cost and low-effort scalability of information extraction to terabytes of text
    - Mímir: Semantic Indexing and search over text, annotations, and ontologies

Track 2: Programming in GATE

- Module 5: GATE Embedded API
    - The GATE component model (LRs, PRs, VRs, GATE Factory)
    - The GATE data model (annotations, documents, corpora)
    - Execution control (controllers, application persistence, compositionality)
- Module 6: Main GATE APIs
    - Advanced JAPE: using Java on the RHS
    - The Ontology API, and Ontology Population
- Module 7: Creating new Resource Types
    - Writing new Processing Resources
    - Writing new Visual Resources
    - Understanding CREOLE configuration
- Module 8: Advanced GATE Embedded
    - GATE and UIMA
    - GATE-based Web Applications

[18] http://gate.ac.uk/conferences/training-modules.html#module13
[19] http://gate.ac.uk/conferences/training-modules.html

- Groovy in GATE

Track 3: Advanced GATE

- Module 9: Ontologies and Semantic Annotation
  - Introduction to Ontologies
  - GATE Ontology Editor
  - GATE Ontology Annotation Tools for Entities and Relations
  - Automatic Semantic Annotation in GATE
  - Measuring Performance
  - Using the Large Knowledge Base gazetteer (LKB)
  - Using MIMIR: the new semantic indexing and search platform
- Module 10: Advanced GATE Applications
  - Customising ANNIE
  - Working with different languages
  - Complex applications, conditional Processing
  - Section-by-section processing
- Module 11: Machine Learning
  - Machine learning and evaluation concepts
  - Using ML in GATE, entity learning hands-on session, relation extraction hands-on session
- Module 12: Opinion Mining
  - Introduction to opinion mining and sentiment analysis
  - Using GATE tools to perform sentiment analysis
  - Machine learning for sentiment analysis hands-on session
  - Future directions for opinion mining

As it can be seen in Table 7, the GATE training curriculum aligns well with the ECULID modules, in terms of developing the skills required for the data analyst, and partially the data manager. The introductory and programming tracks can be used to acquire the necessary basics for dealing with the tools, while the modules from the advanced track, can be used of gain further expertise in the areas of machine learning or opinion mining.

	Data Architect	**Data Manager**	**Data Analyst**	**Data Application Developer**
Introductory Level	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios
Intermediate Level		Module 2: Querying Linked Data	Module 2: Querying Linked Data	Module 2: Querying Linked Data
	Module 3: Providing Linked Data	Module 3: Providing Linked Data		Module 3: Providing Linked Data
			Module 4: Interaction with Linked Data	Module 4: Interaction with Linked Data
		GATE Track 1: Introduction to GATE	GATE Track 1: Introduction to GATE	

		and Text Mining	and Text Mining	
		GATE Track 2: Programming in GATE	GATE Track 2: Programming in GATE	
	GATE Track 3: Module 9: Ontologies and Semantic Annotation	GATE Track 3: Module 9: Ontologies and Semantic Annotation		GATE Track 3: Module 9: Ontologies and Semantic Annotation
			GATE Track 3: Module 10: Advanced GATE Applications	
			GATE Track 3: Module 11: Machine Learning	
			GATE Track 3: Module 12: Opinion Mining	
Advanced Level				Module 5: Creating Linked Data Applications
	Module 6: Scaling up	Module 6: Scaling up		Module 6: Scaling up

Table 7: Skills Alignment with the GATE Training Modules

3.8 Further Training Initiatives

In this section we describe some further training initiatives and materials, related to Linked Data and data science in general.

The Linked Data community website[20] itself provides a collection of reference tutorials and presentations. Unfortunately, it does not offer a training curriculum or a list of topics that are relevant and should be covered. Instead under "Guides and Tutorials"[21] there is a list of relevant documents, tutorials, slides and frequently asked questions. The provided resources are given below:

- Key Reference Documents
  - Design Issues: Linked Data, by Tim Berners-Lee
  - Linked Data: Evolving the Web into a Global Data Space, by Tom Heath and Christian Bizer
- Textual Guides/Tutorials
  - Linked Data: Evolving the Web into a Global Data Space, by Tom Heath and Christian Bizer
  - How to Publish Linked Data on the Web (Tutorial), by Chris Bizer, Richard Cyganiak, Tom Heath (superseded by the Linked Data book, by Tom Heath and Christian Bizer)
  - Introducing Linked Data and the Semantic Web
  - Introducing Graph Data

[20] http://linkeddata.org
[21] http://linkeddata.org/guides-and-tutorials

- Introducing RDF/XML
- Semantic Modelling
- Introducing RDFS and OWL
- Deploying Linked Data using OpenLink Virtuoso
- Querying Semantic Data
- Learn Linked Data - A growing collection of tutorials, essays, links and discussion about Linked Data and related topics.

- Video Tutorials
  - How to Publish Linked Data on the Web tutorial by Tom Heath, Michael Hausenblas, Chris Bizer, Richard Cyganiak, Olaf Hartig, from ISWC2008, Karlsruhe, Germany.
  - The Web, one huge database ... screen-cast by Michael Hausenblas (see also examples and slides).
- Introductory Slide Sets
  - Quick Linked Data Introduction, by Michael Hausenblas
  - An Introduction to Linked Data, by Tom Heath, from the Semantic Web Austin Linked Data tutorials
  - Tutorial on Linked Data - A Practical Introduction, by Michael Hausenblas
  - 30 Minute Guide to RDF and Linked Data, by Ian Davis, from code4lib2009
- Frequently Asked Questions
  - Linked Data FAQ at linkeddata.org
  - Linked Data FAQ (enterprise focus)

As it can be seen, there is no particular structure to the materials and they are not grouped according to topic or application area. Still, we used part of the resources as further reading in some of the modules, while others were helpful in defining the fundamentals.

One further training initiative that should be be mentioned is the one offered by the METANet[22] project. The Multilingual Europe Technology Alliance aims to bring together researchers, commercial technology providers, private and corporate language technology users, language professionals and other information society stakeholders. Its main goal is to organise a network of parties, jointly working on advancing the current state of Language Technology. In this context, the project aims to realise applications that enable automatic translation, multilingual information, knowledge management and content production across all European languages. META-Net training[23] has no explicit training curriculum, however, it covers a variety of text processing and automatic translation topics, including the following:

- Machine learning
- Machine translation
- Statistical learning
- Textual information access
- Text mining and statistical machine translation
- Statistical multilingual analysis
- Natural language processing

[22] http://www.meta-net.eu
[23] http://www.meta-net.eu/meta-research/training

- Probabilities and language models
- Linguistic parsing
- Part-of-speech tagging approaches

These topics do not directly fit into EUCLID's curriculum, however, they are quire relevant for developing the skill of a data analysts and partially also interesting for a data applications developer. Therefore, the materials can be used in conjunction with the module on querying Linked data and interacting with Linked Data, in order to develop a learning path for a trained professional in the field of data processing, interpretation, and analysis. In fact, the courses offered by Cloudera and EMC already include some of the machine learning and automatic text processing topics listed here.

Another collection of related training courses is given by the Data Science Academy (DSA)[24]. The Data Science Academy is a new project from Data Science London, which offers Data Science courses and workshops. Currently, the website provides a list of relevant courses, which are given below:

The Little List of Free Online Data Science Courses[25]

- Introduction to Data Science by Jeff Hammerbacher at University California, Berkeley
- Introduction to Data Science @coursera
- Learning from Data at California Institute of Technology
- Data Science and Analytics at University California, Berkeley School of Information
- An Introduction to Data Science at Syracuse University (pdf)
- Introduction to Data Mining at Massachusetts Institute of Technology
- Introduction to Data Wrangling at the School of Data
- Data Science Course at Columbia University notes by @mathbabe
- An Introduction to Machine Learning @coursera and Stanford University

The courses are not specifically tailored to Linked Data but cover data science topics in general. We have taken some input from the introduction to data science courses. Still currently, we cannot directly align EUCLID's curriculum with the training offered by DSA. However, since this initiative is currently gaining on popularity, we will continue to monitor the developments and report on new curricula in the upcoming deliverables.

Another channel that is worth mentioning is Coursera[26], which provides a collection of official university courses available online. Currently, there are no specific courses on Linked Data. However, training specific in the context of data is already being offered, for example through the Introduction to Data Science course[27]. We will continue to monitor this channel and report on any newly offered training in the area of Linked Data.

[24] http://datascienceacademy.com
[25] http://datascienceacademy.com/free-data-science-courses/
[26] https://www.coursera.org/
[27] https://www.coursera.org/course/datasci

4 Conclusion

With the growing importance and use of Linked Data principles and technologies, there is also an increased demand for trained data practitioners who are able to develop Linked Data-based solutions. The EUCLID project addresses precisely this need by providing an extensive training curriculum that communicates the fundamental background knowledge but also introduces some advanced and expert-level topics.

This deliverable presents the final version of EUCLID's training curriculum. It is based on six main modules, which aim to gradually build up the trainee's knowledge in the filed. The main covered topics include – Introduction to Linked Data and Application Scenarios, Querying Linked Data, Providing Linked Data, Interaction with Linked Data, Creating Linked Data Applications, and Scaling-up. The modules are grouped into three main levels of topics – introductory, advanced and expertise. The curriculum aims to be practice and hands-on oriented, therefore examples, self-assessment questions and demo application are an important part of the presented materials. We conclude the deliverable by identify existing courses and curricula, determining the main topics that they address and identifying overlaps with the EUCLID curriculum. Therefore, we show how they can be used in combination with EUCLID's modules in order to acquire competencies in cross-disciplinary fields in the general area of data science.