Developing a Curriculum of Open Educational Resources for Linked Data

Alexander Mikroyannidis and John Domingue, Knowledge Media Institute, The Open University
{Alexander.Mikroyannidis, John.Domingue}@open.ac.uk
Maria Maleshkova, Karlsruhe Institute of Technology (KIT)
maria.maleshkova@kit.edu
Barry Norton, British Museum
BNorton@britishmuseum.org
Elena Simperl, University of Southampton
E.Simperl@soton.ac.uk

Abstract

The EUCLID project is developing an educational curriculum about Linked Data, supported by multimodal Open Educational Resources (OERs) tailored to the real needs of data practitioners. The EUCLID OERs facilitate professional training for data practitioners, who aim to use Linked Data in their daily work. The EUCLID OERs are implemented as a combination of living learning materials and activities (eBook, online courses, webinars, face-to-face training), produced via a rigorous process and validated by the user community through continuous feedback.

Keywords

Open Educational Resources, Linked Data, Massive Open Online Courses

Introduction

There is a revolution occurring now in higher education, largely driven by the availability of high quality online materials, also known as Open Educational Resources (OERs). OERs can be described as “teaching, learning and research resources that reside in the public domain or have been released under an intellectual property license that permits their free use or repurposing by others depending on which Creative Commons license is used” (Atkins, Brown, & Hammond, 2007). The emergence of OERs has greatly facilitated online education through the use and sharing of open and reusable learning resources on the Web. Learners and educators can now access, download, remix, and republish a wide variety of quality learning materials available through open services provided in the cloud.

The OER initiative has recently culminated in MOOCs (Massive Open Online Courses), which offer large numbers of students the opportunity to study high quality courses with prestigious universities. These initiatives have led to widespread publicity and also strategic dialogue in the higher education sector. The consensus within higher education is that after the Internet-induced revolutions in communication, business, entertainment, media, amongst others, it is now the turn of universities. Exactly where this revolution will lead is not yet known but some radical predictions have been made including the end of the need for university campuses[1].

Linked Data (Berners-Lee, 2006) has established itself as the de facto means for the publication of structured data over the Web, enjoying amazing growth in terms of the number of organizations committing to use its core principles for exposing and interlinking Big Data for

[1] http://www.theguardian.com/education/2012/nov/11/online-free-learning-end-of-university

seamless exchange, integration, and reuse (Bizer, Heath, & Berners-Lee, 2009). More and more ICT ventures offer innovative data management services on top of Linked Data, creating a demand for Data Scientists possessing skills and detailed knowledge in this area. Ensuring the availability of such expertise will prove crucial if businesses are to reap the full benefits of these advanced data management technologies, and the know-how accumulated over the past years by researchers, technology enthusiasts and early adopters.

The European project EUCLID[2] contributes to this goal by developing a comprehensive educational curriculum, supported by multimodal OERs and highly visible eLearning distribution channels, tailored to the real needs of data practitioners. The EUCLID curriculum focuses on techniques and software to integrate, query, and visualize Linked Data, as core areas in which practitioners state to require most assistance. A significant part of the learning materials produced in the project consists of examples referring to real-world data sets and application scenarios, code snippets and demos that developers can run on their machines, as well as best practices and how-tos.

The EUCLID approach

The EUCLID educational curriculum consists of a series of modules, each containing multi-format OERs, such as presentations, webinars, screencasts, exercises, eBook chapters, and online courses. These learning materials complement each other and are connected to deliver a comprehensive and concise training programme to the community. Learners are guided through these materials by following learning pathways, which are sequences of learning resources structured appropriately for achieving specific learning goals. Different types of eLearning distribution channels are targeted by each type of learning materials, including Apple and Android tablets, Amazon Kindles, as well as standard web browsers (see Figure 1). The EUCLID learning materials are available for free on the project web site, as well as on Apple's iBook Store[3] as an interactive iBook for use on the iPad and MacOS. All the materials are made available under a Creative Commons Attribution 3.0 Unported License[4].

Instead of mock Linked Data examples, we use in our learning materials and exercises a collection of datasets and tools that are deployed and used in real life. In particular, we use a number of large datasets including, for example, the MusicBrainz dataset, which contains 100Ms of triples. Our collection of tools includes Seevl, Sesame, Open Refine and GateCloud, all of which are used in real-life contexts. We also showcase scalable solutions, based upon industrial-strength repositories and automatic translations, e.g. by using the W3C standard R2RML for generating RDF from large data contained in standard databases.

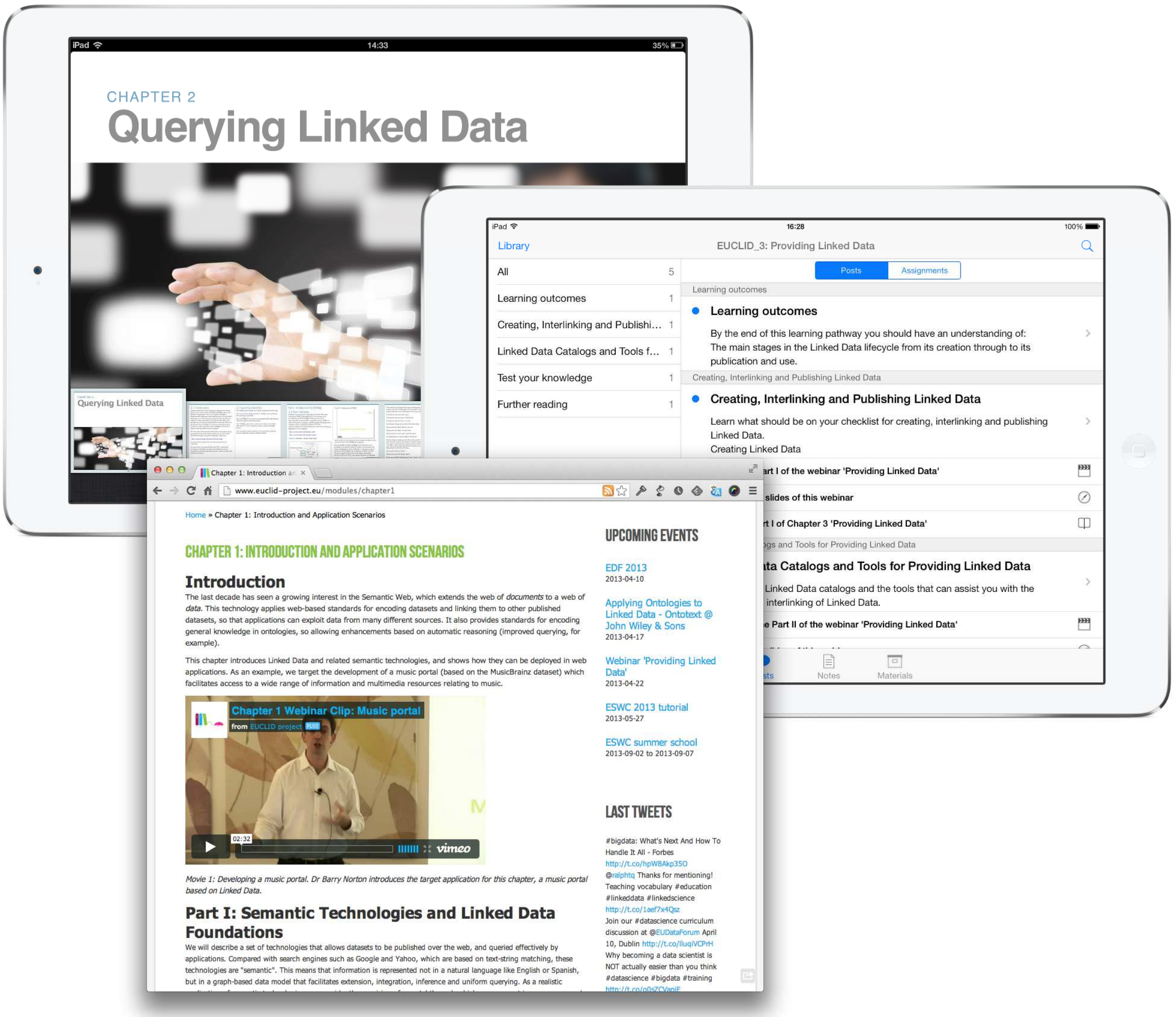
Additionally, EUCLID has a strong focus on the community and encourages community engagement in the production of OERs through, for example, collecting user feedback via our webinars, Twitter, LinkedIn, and more. EUCLID combines online and real-world presence, and attempts to integrate with on-going activities in each sphere such as mailing lists and wikis. The project engages with the Linked Data community, both practitioners and academics, by

[2] http://www.euclid-project.eu

[3] http://bit.ly/using-linked-data-effectively

[4] http://creativecommons.org/licenses/by/3.0

collecting user requirements as well as feedback to the OERs so that they can be tailored to what the learner really needs.


iPad
14:33
35%
CHAPTER 2
Querying Linked Data
Querying Linked Data
iPad
16:28
100%
Library
EUCLID_3: Providing Linked Data
All 5
Learning outcomes 1
Creating, Interlinking and Publishi… 1
Linked Data Catalogs and Tools f… 1
Test your knowledge 1
Further reading 1
Posts
Assignments
Notes
Materials
Chapter 1: Introduction a
www.euclid-project.eu/modules/chapter1
Home » Chapter 1: Introduction and Application Scenarios
Chapter 1 Webinar Clip: Music portal
from EUCLID project PLUS
02:32
vimeo

Learning outcomes

Learning outcomes

By the end of this learning pathway you should have an understanding of:
The main stages in the Linked Data lifecycle from its creation through to its publication and use.

Creating, Interlinking and Publishing Linked Data

Creating, Interlinking and Publishing Linked Data

Learn what should be on your checklist for creating, interlinking and publishing Linked Data.
Creating Linked Data

art I of the webinar 'Providing Linked Data'

slides of this webinar

rt I of Chapter 3 'Providing Linked Data'

gs and Tools for Providing Linked Data

ta Catalogs and Tools for Providing Linked Data

Linked Data catalogs and the tools that can assist you with the interlinking of Linked Data.

e Part II of the webinar 'Providing Linked Data'

CHAPTER 1: INTRODUCTION AND APPLICATION SCENARIOS

Introduction

The last decade has seen a growing interest in the Semantic Web, which extends the web of *documents* to a web of *data*. This technology applies web-based standards for encoding datasets and linking them to other published datasets, so that applications can exploit data from many different sources. It also provides standards for encoding general knowledge in ontologies, so allowing enhancements based on automatic reasoning (improved querying, for example).

This chapter introduces Linked Data and related semantic technologies, and shows how they can be deployed in web applications. As an example, we target the development of a music portal (based on the MusicBrainz dataset) which facilitates access to a wide range of information and multimedia resources relating to music.

Movie 1: Developing a music portal. Dr Barry Norton introduces the target application for this chapter, a music portal based on Linked Data.

Part I: Semantic Technologies and Linked Data Foundations

We will describe a set of technologies that allows datasets to be published over the web, and queried effectively by applications. Compared with search engines such as Google and Yahoo, which are based on text-string matching, these technologies are "semantic". This means that information is represented not in a natural language like English or Spanish, but in a graph-based data model that facilitates extension, integration, inference and uniform querying. As a realistic

UPCOMING EVENTS

EDF 2013
2013-04-10

Applying Ontologies to Linked Data - Ontotext @ John Wiley & Sons
2013-04-17

Webinar 'Providing Linked Data'
2013-04-22

ESWC 2013 tutorial
2013-05-27

ESWC summer school
2013-09-02 to 2013-09-07

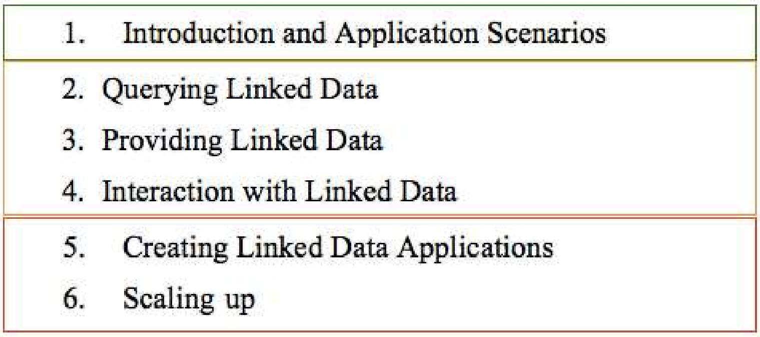
LAST TWEETS

#bigdata: What's Next And How To Handle It All - Forbes http://t.co/hpW8Akp35O
@ralphtq Thanks for mentioning! Teaching vocabulary #education #linkeddata #linkedscience http://t.co/1aef7x4Qsz
Join our #datascience curriculum discussion at @EUDataForum April 10, Dublin http://t.co/IluqiVCPrH
Why becoming a data scientist is NOT actually easier than you think #datascience #bigdata #training http://t.co/q0sZCVapiF

Figure 1. A selection of EUCLID learning materials in different formats and platforms, i.e. eBooks and online courses for the web and the iPad.

The EUCLID curriculum

The EUCLID curriculum has been designed to gradually build up the learner's knowledge. It enables learners with previous knowledge on a specific area of interest to only briefly go over the introductory materials and directly dig into one of the more advanced modules. As shown in Figure 2, the EUCLID curriculum is organized in the order of 3 expertise levels (top: introductory, middle: advanced, bottom: expertise). It is composed of 6 modules that cover all the major aspects of the Linked Data consumption lifecycle.



1. Introduction and Application Scenarios
2. Querying Linked Data
3. Providing Linked Data
4. Interaction with Linked Data
5. Creating Linked Data Applications
6. Scaling up

Figure 2. The EUCLID curriculum

The 6 EUCLID modules have been structured to cover the following range of topics:

- *Module 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 dataset), 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.
- *Module 2: Querying Linked Data*. 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.
- *Module 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.
- *Module 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 used. 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.

- *Module 5: Creating Linked Data Applications*. This module gives details on technologies and approaches towards exploiting Linked Data by building bespoke 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 mash-ups.
- *Module 6: Scaling up*. This module addresses the main issues of Linked Data and scalability. In particular, it provides gives details on approaches and technologies for clustering, distributing, sharing, and caching data. Furthermore, it addresses the means for publishing data trough could deployment and the relationship between Big Data and Linked Data, exploring how some of the solutions can be transferred in the context of Linked Data.

In an effort to provide high-quality training, suitable for the data practitioner's needs, the EUCLID curriculum has been through several revisions on structure, arrangement and content after presenting it to a number of experts and gathering their feedback. As a result of these revisions, the curriculum was refined and developed in more detail in order to include a number of expected outcome competencies, as well as a variety of exercises and examples. The content of the EUCLID modules has been redesigned to be better aligned and support a smoother process of skills built-up and development. While having an individual objective, each module contributes to further developing the skills and knowledge gained by the previous one thus aiding to acquiring an overall understanding and expertise in the field. As mentioned before, the curriculum is constantly updated based on feedback from the community.

The EUCLID OER production process

The OER production process defines the sequence of steps for the production of the EUCLID learning materials. Initially, 3 basic steps were planned in order to create each module and its exercises (see Figure 3). Firstly, following the curriculum, the draft of the training material would be created, which includes slides for a webinar, as well as HTML content for online distribution. Secondly, feedback on the drafts would be gathered and analysed. Finally, based on the comments and feedback, each module would be refined before delivering an eBook encompassing all the training materials, which include written documents, examples, presentation slides, as well as the video recording of the webinar.

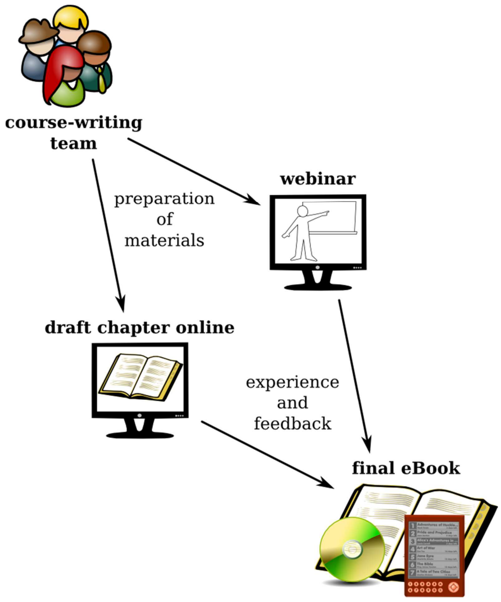


Figure 3. The initial OER production process

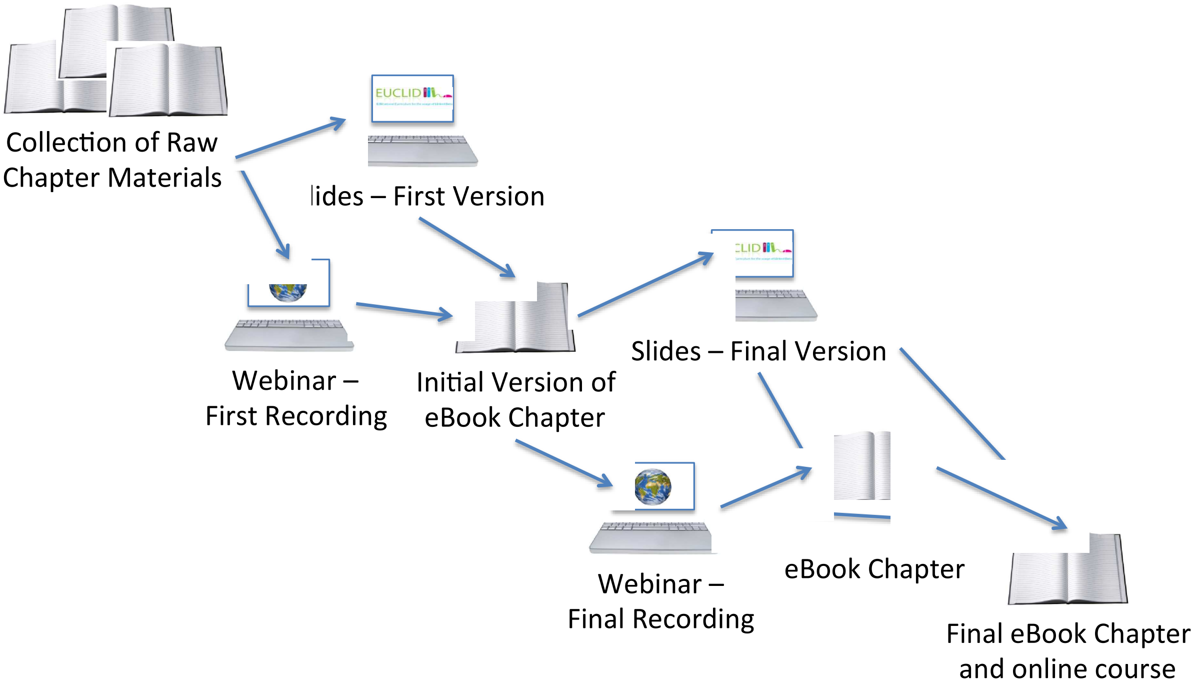

Collection of Raw Chapter Materials
lides – First Version
Webinar – First Recording
Initial Version of eBook Chapter
Slides – Final Version
Webinar – Final Recording
eBook Chapter
Final eBook Chapter and online course

Figure 4. The revised OER production process

During the production of the first EUCLID module, this process was further refined and elaborated to include some intermediate steps (see Figure 4). One thing that became obvious was the instrumental role of the preparation and delivery of the webinar in the production process. The webinar was therefore produced in two stages. First, an internal webinar was held in order to collect feedback from project partners about its content and structure. The learning materials were revised through collecting comments and feedback from the internal broadcasting of the first webinar and the publication of the first version of the eBook chapter.

Subsequently, a second version of the webinar was produced, this time publically broadcasted. Based on the community feedback received from the broadcasting of the second webinar, the structure and content of the module were finalised and the eBook chapter was produced from all the finalised content. It was also decided that additional material in the form of an online course would accompany the final eBook chapter and would be part of the training programme offered to the community. This process has been applied for the production of all EUCLID modules.

Best practices for the design and delivery of Linked Data OERs

The design and implementation of the OER production process has provided us with a valuable insight into the various challenges associated with the design and delivery of learning materials specifically for Linked Data. We have thus distilled our experiences and lessons learned into a set of best practices, which is outlined in the next 2 sections.

Best practices for the design of Linked Data OERs

1. **Industrial Relevance –** our curriculum takes into account the needs of industry related to Open Data and Linked Data. Future work aims to automatically mine and analyse relevant job adverts to gain desired competencies for the sector. This is supported by the following best practice.
2. **Team Curriculum Design –** where the team is composed of a number of roles to fully capture industrial, academic and pedagogical requirements. Our team comprises of industrial partners (Ontotext, FluidOps), who have extensive experience with professional training, industrial requirements and scalable tools, academic partners (KIT, STI International), who have research expertise in Linked Data and pedagogical experts (The Open University).
3. **External Collaboration –** to gain world-class curriculum expertise where necessary and to facilitate course delivery and dissemination.
4. **Explicit learning goals –** to which all learning materials (slides, webinars, eBook chapters) are developed. Learners are guided through the learning goals by learning pathways – a sequence of learning resources to achieve a learning goal.
5. **Show realistic solutions –** rather than mock examples we utilize systems that are deployed and used for real.
6. **Use real data –** we use a number of large datasets including for example, the MusicBrainz dataset that contains 100Ms of triples.
7. **Use real tools –** our collection of tools are used in real life, including for example Seevl, Sesame, Open Refine and GateCloud.
8. **Show scalable solutions –** based upon industrial-strength repositories and automatic translations, for example using the W3C standard R2RML for generating RDF from large data contained in standard databases.
9. **Eating our own dog food –** we monitor communication and engagement with the Linked Data community through W3C email lists, in the social network channels LinkedIn and Twitter, as well as content dissemination channels such as Vimeo and SlideShare. We transform the monitoring results into RDF and make these available at a SPARQL endpoint. In this respect we use Linked Data to support Learning Analytics.

Best practices for the delivery of Linked Data OERs

1. **Open to Format** – our learning materials are available in a variety of formats including: HTML, iBook (iPad and MacOS), ePUB (Android tablets), MOBI (Amazon Kindle).
2. **Addressability** – every concept in our curriculum is URI-identified so that HTML and RDF(a) machine-readable content is available.
3. **Integrated** – to ease navigation for learners the main textual content, relevant webinar clips, screencasts and interactive components are placed into one coherent space.
4. **High Quality** – we have a formalised process where all materials go through several iterations to ensure quality. For example, for each module we run both a practice and a full webinars facilitating critique and commentary.
5. **Self-testing and reflection** – in every module we include inline quizzes and exercises formulated against learning goals enabling students to self-monitor their progress.

Conclusion

The EUCLID project has established a rigorous process for the production and delivery of OERs about Linked Data. This process defines a series of iterations in the production of learning materials, with multiple revisions from internal and external stakeholders, in order to ensure high quality in the produced materials. Based on our experiences and lessons learned in designing and implementing the production process, we have also established a set of best practices for the design and delivery of OERs specifically for Linked Data.

One of our main goals is to reach out to the community in as many ways as possible, in order to engage it and acquire its feedback. For this purpose, considerable effort has been put to delivering the learning materials in a variety of formats and for different purposes. The EUCLID learning materials can be accessed from a wide range of platforms, both from desktop/laptop computers, as well as from different mobile devices. With the learning materials reaching a constantly growing community, it is expected that there will be more comments and feedback received, which we will continuously monitor in order to improve the quality of the EUCLID training programme.

References

Atkins, Daniel E., Brown, John Seely, & Hammond, Allen L. (2007). A Review of the Open Educational Resources (OER) Movement: Achievements, Challenges, and New Opportunities (pp. 4): The William and Flora Hewlett Foundation.

Berners-Lee, T. (2006). Linked Data - Design Issues. from http://www.w3.org/DesignIssues/LinkedData.html

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data—The Story So Far. *International Journal on Semantic Web and Information Systems, 5*(3), 1–22.