ATHENA
Access to cultural heritage networks across Europe

ECP-2007-DILI-517005

ATHENA

Specific tools to be used for conversion and adaptation of proprietary museum data

Deliverable number	*D3.6*
Dissemination level	*Public*
Delivery date	*30 April 2011*
Status	*Final*
Author(s)	*Gordon McKenna, Collections Trust (UK), Roxanne Wyns, Royal Museums of Art and History (BE)*

e*Content*plus

This project is funded under the eContentplus programme1,
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

[1] OJ L 79, 24.3.2005, p. 1.

Table of Contents

1. Introduction

1.1 The purpose and results of ATHENA work package 3

As this is the final deliverable of work package 3 of the ATHENA project (WP 3) is perhaps a good time to present a short review of its work. WP 3 was tasked with:

1. Reviewing the different standards in use by museums;
2. Facilitating the mapping of those standards to a common metadata standard;
3. Assessing the requirements for the persistent identification of digital objects and collections;
4. Producing tools to support the conversion of museums' data into the common harvesting format for ingestion into the main Europeana service.

Task 1 – This began with a survey of technical and metadata standards being used by the content providers listed in the ATHENA project's *Description of Work*. The survey also included questions on the use of terminologies, and the IPR situation with regard to the collections. These were used by WP 4 and WP 5 respectively for their work. The survey led to a report and a set of best practise recommendations. The deliverables of this task were:

- *D3.1 – Report on existing standards applied by European museums*.[2]
- *D3.2 – Recommendations and best practice report regarding the application of standards, including recommendations for a harvesting format and fact sheets for dissemination*.[3]

Task 2 – The analysis of metadata standards use revealed by Task 1 led to the ATHENA project seeking to develop and deploy a XML harvesting schema suitable for providing rich museum metadata to portals. Another result was the decision to create a schema based on existing schemas and standards – ***LIDO* (Light Information Describing Objects)**. The deliverable for this task was:

- *D3.3 – Specification for conversion tools*.[4] (a description of how LIDO was created, and a mapping document 'tool')

However in addition to this deliverable WP3 also produced two other 'deliverables' in support of the specification:

- *The LIDO XML schema*.[5] (A description of the schema)
- *A technical description of LIDO*.[6] (The schema itself)

After this initial publication further work was carried out on the schema and this has resulted in updates to the schema and its documentation:

- *LIDO v1.0 Specification Document.*[7] (documentation)

[2] See: **http://www.athenaeurope.org/getFile.php?id=396**
[3] See: **http://www.athenaeurope.org/getFile.php?id=538**
[4] See: **http://www.athenaeurope.org/getFile.php?id=539**
[5] See: **http://www.athenaeurope.org/getFile.php?id=535**
[6] See: **http://www.athenaeurope.org/getFile.php?id=536**

- *LIDO v1.0 XML Schema Definition*.[8] (schema)

These were published within the framework of Data Harvesting and Interchange Working Group of CIDOC[9].

LIDO is implemented in the *ATHENA Project Ingestion Server* for the mapping and conversion of non-standard metadata, and the delivery of ESE (Europeana Semantic Elements) metadata to Europeana.

Task 3 – Using and maintaining persistent identifiers (PIDs) is an important 'tool' for consistent delivery of data to Europeana. The task began with a short survey to look at the current use of PIDs. This, together with research into the available standards and systems for PIDs, led to the definition of a landscape, and to best practise advice. This was followed by similar work on best practise for the policies and technical infrastructural needs for supporting PIDs. The deliverables of this task were:

- *D3.4 – Assessment of requirements for persistent identification of objects, collections and institutions*.[10]
- *D3.5 – Technical and policy infrastructure to support persistent identifiers*.[11] (including a 'white paper' on PIDs)

Task 4 – Tools were created and integrated throughout most of the project work packages. Existing tools, from non-ATHENA sources, were also used. This deliverable looks these, and adds a final tool in the form of guidance for the process workflow of providing material to Europeana.

1.2 Overview of the deliverable

This deliverable is defined in the ATHENA *Description of Work* as of the type 'Demonstrator', and covers the 'tools' produced during the ATHENA project. There is no external definition of exactly what a tool is the context of ATHENA, therefore we define them as: software systems (with helpdesk support); metadata standards; training materials; and guidelines documents. All of these are covered in the deliverable.

The deliverable is in two parts:

- An overview of the tools created, and used, during the project which have helped partners to provide their material to Europeana.
- A guide, based on partners' experiences during the project, which gives users an overview on how to get the most from their content and metadata when providing it to Europeana.

These tools were used in the context of the ATHENA project; however most of them have a general application. The *ATHENA Project Ingestion Server* is being applied in other current Europeana

[7] See: **http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf**
[8] See: **http://www.lido-schema.org/schema/v1.0/lido-v1.0.xsd**
[9] See: **http://www.lido-schema.org**
[10] See: **http://www.athenaeurope.org/getFile.php?id=725**
[11] See: **http://www.athenaeurope.org/getFile.php?id=772**

Group projects, e.g. *Judaica Europeana* and *MIMO*. It will also be maintained and used in the new projects like DCA and *Linked Heritage* (starting April 2011).

2. Tools created and used during the ATHENA project

There were many tools used by partners to submit their material to Europeana during the ATHENA project. These tools were created by:

- Partners in ATHENA Work Package 3;
- Partners in other work packages (i.e. WP 4, WP 5, and WP 7);
- Other EC-funded projects contemporary with ATHENA (including Europeana itself);
- Other EC-funded projects prior to ATHENA.

The authors of this deliverable acknowledge the work of the creators of the tools, and also would like to thank the partners of WP 2 who worked on turning some of the more 'dry' deliverables of the ATHENA project into attractive publications, available both electronically and in printed form.

Here we look at the most used tools during the project.

2.1 Describing tools

In order to be consistent throughout all the deliverables of this work package we have decided again to use the metadata scheme developed for deliverable D3.1. As before, we describe each tool in a Dublin Core (DC) derived format. 9 out of the 15 DC elements are used.

These elements are:

Title	The name under which the tool is known.
Creator	The name of the organisation which originally created the tool and/or authors.
Publisher	The name of the organisation that makes tool available.
Date	The date on which the tool was published or made available.
Identifier	A number or other identifier under which a tool is published and/or a URL which points to the tool online.
Rights	What the use rights of the tool are.
Description	A textual description explaining the tool and its usage.
Subject	Keywords that identify the tool.
Relation	Other tools that this one relates to, and associated websites.

The descriptions are aimed at a general reader in a cultural heritage organisation. They are especially for a museum, with rich information about its collections, planning to provide it to Europeana. The purpose is to allow the reader to have an easy reference to the range of tools they might need to use in one place.

2.2 Standards guidance

ATHENA's work on technical standards was an early part of Work Package 3's activities. It began with a survey on the metadata and technical standards used by ATHENA partners. This led to the publication of:

Title	***Digitisation: Standards Landscape for European Museums, Archives, Libraries***
Creator	McKenna, Gordon and De Loof, Chris (text)
Publisher	ATHENA Project
Date	2009
Identifier	**http://www.athenaeurope.org/getFile.php?id=435**
Rights	Creative Commons (CC-BY-NC-SA)
Description	Contains basic descriptions of standards in the areas of: ***Information schemes (metadata)*** used for**:** • Museum specific; • Archive specific; • Library specific; • Historic environment specific; • General heritage; • Resource discovery; • Document encoding. ***Multimedia formats*** for: • Text; • Image; • Audio; • Video; • Virtual reality; • Vector graphics. Users gain basic information about the wide range of standards available for use in digitisation.
Subject	archive description; bibliographic description; documentation (historic environment); documentation (museum); documentation (visual culture); collection description; resource discovery; description (cultural object); image format; sound format; video format; virtual reality; animated vector graphics; raster graphics; vector graphics; document encoding; document rendering; document structure; page layout language; character encoding; non-Western scripts; data transmission; digital library; file transfer; harvesting protocol; hypertext transfer; interactivity; multimedia; panoramas; query language; relational databases; search and retrieval protocol.
Relation	**http://www.athenaeurope.org/getFile.php?id=396** (ATHENA deliverable D3.1)
	http://www.minervaplus.ru/publish/standards_landscape.pdf (Russian)

The survey work was followed by a set of recommendations and best practise guidelines:

Title	***ATHENA D3.2 – Recommendations and best practice report regarding the application of standards, including recommendations for a harvesting format and fact sheets for dissemination***
Creator	McKenna, Gordon and De Loof, Chris
Publisher	ATHENA Project
Date	2009
Identifier	**http://www.athenaeurope.org/getFile.php?id=538**
Rights	Creative Commons (CC-BY-NC-SA)
Description	Contains advice for using metadata and technical content standards in three 'use environments': • Master (digital content) or Collections Management (metadata); • Service; • Discovery. The recommendations are summarised in two fact sheets. The metadata harvesting standard recommended for use by museums in the Service environment is LIDO (Light Information Describing Objects). This was developed during the ATHENA project.
Subject	digital content use environment; metadata use environment; best practice (standards)
Relation	**http://www.minervaeurope.org/publications/MINERVA%20TG%202.0.pdf** (Minerva Guidelines)

The *Minerva Guidelines* were especially mentioned:

Title	***Technical Guidelines for Digital Cultural Content Creation Programmes: Version 2.0***
Creator	Fernie, Kate; De Francesco, Giuliana; and Dawson, David (eds.)
Publisher	Minerva Project
Date	2008
Identifier	**http://www.minervaeurope.org/publications/MINERVA%20TG%202.0.pdf**
Rights	MINERVA eC Project
Description	Contains: • ***Projects and Planning***; • ***Preparing for the digitisation process***; • ***Storage and Management of the Digital Master Material***; • ***Metadata, standards and resource discovery***; • ***Publishing on the Web***; • ***Delivery formats***;

	• ***Reuse and Re-purposing***; • ***Intellectual Property Rights, Copyright, Licensing and Sustainability***.
Subject	best practice (standards); best practice (projects); best practice (digitisation)
Relation	**http://www.minervaeurope.org/structure/workinggroups/servprov/documents/technicalguidelines1_0.pdf** (Version 1 – English)
	http://www.digital-heritage.at/upload/technicalguidelines.pdf (Version 1 – German)
	http://www.kunstenenerfgoed.be/ake/view/nl/1601431-Technische+Richtlijnen+voor+Programmas+voor+de+Creatie+van+Digitale+Culturele+Content.html (Version 1 – Dutch)
	http://www.minervaeurope.org/structure/workinggroups/servprov/documents/techguid1_0-f.pdf (Version 1 – French)
	http://www.minervaeurope.org/structure/workinggroups/servprov/documents/technicalguidelinesita1_8.pdf (Version 1 – Italian)
	http://digitization.hpclab.ceid.upatras.gr/Odhgos_kalwn_praktikwn1.0.pdf (Version 1 – Greek)

This document is **highly recommended** to any organisation already carrying out a digitisation project and especially to those who are considering beginning one.

2.3 LIDO XML harvesting schema

LIDO (Light Information Describing Objects) is an XML harvesting schema that was developed to meet the needs of museum community to represent its, potentially, rich information in portals.

Title	***Lightweight Information Describing Objects (LIDO): The international harvesting standard for museums***
Creator	McKenna, Gordon; Rohde-Enslin, Stefan and Stein, Regine (text)
Publisher	ATHENA Project
Date	2011
Identifier	**http://www.athenaeurope.org/getFile.php?id=786**
Rights	Creative Commons (CC-BY-NC-SA)
Description	A user-friendly introduction to the LIDO XML harvesting schema. Contains: • ***Why is LIDO needed?*** • ***LIDO's background***; • ***LIDO's home***; • ***When to use LIDO***; • ***Basic design principles***.
Subject	XML; metadata (museum); metadata harvesting
Relation	**http://www.lido-schema.org/schema/v1.0/lido-v1.0.xsd** (XML schema)

	http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf (specification document)
	http://www.lido-schema.org (website)

The relations listed point to CIDOC-maintained version of LIDO. They give much more detailed technical documentation, and the schema itself.

2.4 ATHENA Project Ingestion Server

This is the major tool developed within the framework of the ATHENA project. It is: a XML metadata ingestion tool; mapping tool; and an OAI-PMH repository. It was created especially for the ATHENA project. It met the needs to store rich metadata about museum objects, and submit such information to Europeana.

Title	ATHENA Project Ingestion Server
Creator	National Technical University of Athens (NTUA)
Publisher	ATHENA Project
Date	2009–
Identifier	**http://athena.image.ntua.gr/athena** [ATHENA project partners only]
Rights	[Open source developed]
Description	With this open source developed tool organisations with their own metadata, in XML form, can: • **Create and manage organisational and user profiles.** Organisations can have sub-organisations. Users can be assigned to any organisation and have different permissions for viewing and editing. • **Import their metadata in any XML format.** The system checks that the imported metadata is valid XML. (LIDO XML itself can be imported). Import methods are: HTTP upload, FTP (NTUA and remote servers), and OAI-PMH harvesting. • **Map their XML metadata to LIDO.** This is most simply done by a one-to-one element mapping, using a 'drag and drop' mechanic. However more complex mappings can be carried out like: conditional mapping; concatenation, and assignment of 'constant values' where data is absent from the imported metadata. Users of the system can see and have access to the XML transformation control documents (XSLTs). A preview of how records will appear in Europeana is also available. • **Make their transformed data available to Europeana.** When Europeana has been given permission it harvests the data as valid ESE records, using OAI-PMH, and eventually a provider's records will become visible in the Europeana portal.
Subject	XML; data mapping; OAI-PMH; LIDO (schema); ESE (schema)

Relation	**http://www.lido-schema.org/schema/v1.0/lido-v1.0.xsd** (LIDO XML schema)
	http://version1.europeana.eu/web/guest/technical-requirements (link to latest ESE XML schema)

For information on the ATHENA help desk for the System see below.

2.5 ATHENA System training materials and help desk

In January 2010 the ATHENA project gave two workshops in Rome and Berlin. The aim was to train partners in the use of the *ATHENA Project Ingestion Server* and in LIDO. Updates to LIDO and the System were presented at a plenary meeting of the project in Ljubljana in June. Details can be found at:

Title	***ATHENA Training***
Creator	ATHENA Project (various contributors)
Publisher	ATHENA Project
Date	2010
Identifier	**http://www.athenaeurope.org/index.php?en/159/training**
Rights	ATHENA Project
Description	Contains: • ***Introduction to LIDO*** (PDF and VIDEO) • ***Using LIDO*** (PDF and VIDEO x2) • ***ATHENA LIDO mapping worksheet*** (Microsoft Word) • ***Ljubljana*** (VIDEO x2)
Subject	LIDO; ATHENA Project Ingestion Server
Relation	**http://athena.image.ntua.gr/athena** **(**ATHENA Project Ingestion Server)
	http://www.athenaeurope.org/getFile.php?id=565 (*Introduction to LIDO*, PDF)
	http://www.athenaeurope.org/index.php?en/162/training-workshop-rome-18-january-2010-regine-stein-introduction-to-lido (*Introduction to LIDO*, VIDEO)
	http://www.athenaeurope.org/getFile.php?id=560 (*Using LIDO*, PDF)
	http://www.athenaeurope.org/index.php?en/163/training-workshop-rome-18-january-2010-gordon-mckenna-using-lido (*Using LIDO* [1], VIDEO)
	http://www.athenaeurope.org/index.php?en/165/training-workshop-rome-18-january-2010-gordon-mckenna-using-lido-2 (*Using LIDO* [2], VIDEO)
	http://www.athenaeurope.org/getFile.php?id=561 (*ATHENA LIDO mapping worksheet*, WORD)
	http://www.athenaeurope.org/getFile.php?id=654 (*Ljubljana* [1], VIDEO)

	http://www.athenaeurope.org/getFile.php?id=655 (*Ljubljana* [2], VIDEO)

All these materials are available for anyone to use. They act as tools enabling the use of the ATHENA System, and so lead to the conversion of proprietary data into a standard format (LIDO). This in turn is converted to ESE for harvesting by Europeana. However having access to the ATHENA System is essential to get the maximum benefit.

The use of the System during the ATHENA project was supported by the *ATHENA Help Desk*. This provided *ad hoc* support mainly through an e-mail list system, but this was supplemented by telephone (or Skype) help, where needed.

2.6 Terminologies

The use of standard data, terminologies, is important for the interoperability of metadata, especially in aggregation environments like Europeana. In the ATHENA project this issue was the responsibility of WP 4 – *Integration of existing data structure into the EDL*). Its aims were to:

- Analyse and compare existing dictionaries, terminologies, thesauri, classifications, and taxonomies used by museums in a cross-domain perspective.
- Analyse and compare existing, or possible, multilingual tools (thesauri, cross-language retrieval tools, and technical solutions) for access to resources available in museums and other cultural heritage institutions;
- Make recommendations to data providers aimed at facilitating the semantic integration of their content into Europeana.

The deliverables created by this work are integrated into:

Title	***ATHENA WP4 Wiki***
Creator	Leroi, Marie-Véronique and Holland, Johann (et al)
Publisher	ATHENA Project
Date	2009-
Identifier	**http://www.athenaeurope.org/athenawiki**
Rights	ATHENA Project
Description	Aimed at people working in European museums, experts or non-experts in information engineering, and linguistics, who have an interest in terminology and multilingualism, or just want to have general information on the topic. It includes: • Step by step recommendations for making digital resources available and exploitable in Europeana; • A guide to terminology management; • Information about SKOS, a standard format for terminologies, and some material on how to use it; • Access the *ATHENA Thesaurus* and enrich it. This thesaurus is an

	experiment in the production of a skosified multilingual terminology. It is based on existing resources (*Michael Terminology lists - Subjects*; *PICO thesaurus*; and the unpublished *RMCA Keywords*); • Inventory of terminology resources; • The context of ATHENA WP4 and its activities.
Subject	lexicon; dictionary; folksonomy; glossary; classification; taxonomy; thesaurus; controlled vocabulary; terminology; ontology; SKOS (Simple Knowledge Organisation System); skosification
Relation	**http://www.athenaeurope.org/getFile.php?id=398** (ATHENA deliverable D4.1)
	http://www.athenaeurope.org/getFile.php?id=684 (ATHENA deliverable D4.2)
	http://www.michael-culture.org/software/lists.zip (Michael Terminology lists - Subjects)
	http://www.culturaitalia.it/pico/thesaurus/4.1/thesaurus_4.1.0.skos.xml (PICO thesaurus)

2.7 Geographical information

One of the tasks of WP 7 was *Guidelines for geographic location description of digital cultural content*. The scale of the knowledge and use 'gap' was indicated by the results of the survey carried out by WP 3. This showed use of geographical information in organisations was:

- Geographic names terminologies – 44.4%
- Geographic co-ordinates – 8.9%

The use of geographic name terminologies is reasonably high, but the interoperability could be questioned. Only 45% of those using terminologies were using published standards. Also most the published terminologies used national and were monolingual.

The use of geographical co-ordinate information is low in the organisations sampled. This probably reflects the limited use that is made of geographical information by cultural organisations, particularly museums, in displaying their collections online.

Therefore there is an obvious need for basic guidelines on the geographic information: standards to be used, and how this information can be used in the online environment. These were produced and are summarised in:

Title	***Digital Cultural Content: Guidelines for geographic information***
Creator	Zakrajsek, Franc J (text)
Publisher	ATHENA Project
Date	2011
Identifier	**http://www.athenaeurope.org/getFile.php?id=787**
Rights	Creative Commons (CC-BY-NC-SA)
Description	A user-friendly introduction to geographical information for those working in cultural heritage. Contains: • ***Basic terms***; • ***What is a Geographic Information System?*** • ***Standards***; • ***Possible cases of use***.
Subject	geographical information system (GIS)
Relation	**http://www.isotc211.org/Outreach/ISO_TC_211_Standards_Guide.pdf** (ISO/TC 211 Geographic information/Geomatics Standards Guide)
	http://www.opengeospatial.org/standards (OpenGIS® standards)
	http://inspire.jrc.ec.europa.eu (INSPIRE Directive)

2.8 Persistent Identifiers (PIDs)

There is a need for cultural heritage information and content to be persistently available online. At any one time a significant number of URLs on the Europeana portal are broken. This is not solely a Europeana issue. Any portal or website pointing to resources outside its direct control experiences this difficulty. It is even possible that links that are under control of an organisation become broken in the organisation's own service!

One way for this situation to be 'solved' is for the provider to use and manage persistent identifiers (PIDs) for its physical and digital objects. The assumption of the ATHENA project was that there is a lack of knowledge in the sector as a whole. Therefore work was carried out in the project which addressed this need and the tool below gives basic information. Fuller information can be found in deliverables of the project given as relations to it.

Title	***Persistent Identifiers (PIDs): Recommendations for institutions***
Creator	McKenna, Gordon and Wyns, Roxanne
Publisher	ATHENA Project
Date	2011
Identifier	**http://www.athenaeurope.org/getFile.php?id=779**

Rights	Creative Commons (CC-BY-NC-SA)
Description	A user-friendly introduction to persistent identifiers (PIDs) for those working in cultural heritage. Contains: • Persistent identifiers: A briefing note; • Persistent identifier policy in context; • Standards landscape; • Managing organisations; • Persistent identifier systems. The standards landscape looks at PIDs for: • Physical objects in museums; • Digital objects; • Collections in museums; • Institutions; • PID Services. It is derived from two deliverables of the ATHENA project.
Subject	persistent identifier; URI (Uniform Resource Identifier); URL (Uniform Resource Locator); URN (Uniform Resource Name); PURL (Persistent URL); Handle System; DOI (Digital Object Identifier); OpenURL; ARK (Archival Resource Key); ISIL (International Standard Identifier for Libraries and Related Organizations); MDA Code.
Relation	**http://www.athenaeurope.org/getFile.php?id=725** (ATHENA deliverable D3.4)
	http://www.athenaeurope.org/getFile.php?id=772 (ATHENA deliverable D3.5)

2.9 Intellectual Property Rights

Knowledge of the rights, and particularly Intellectual Property Rights (IPR), is an important prerequisite before attempting to make content available to Europeana. The importance of this recognised by Europeana and they give a set of *Europeana Rights Guidelines* which sets out their requirements which find instantiation in the ESE metadata schema, and is an important part of a item record on a Europeana portal.

WP 6 (*Analysis of IPR issues and definition of possible solutions*) carried out work in this area. That work can be summarised in the online:

Title	***Step-by-step IPR Guide***
Creator	Dierickx, Barbara and Tsolis, Dimitrios
Publisher	ATHENA Project
Date	2010
Identifier	**http://devel.silktech.gr/athenaeurope_ipr/**

Rights	Creative Commons (CC-BY-NC-SA)
Description	Aims to address the problems which may occur when an organisation is trying to bring its collection to the Web. It proposes a straightforward method of clearing rights in order to achieve the legal basis for the use needed. The steps are: • Identification of the work; • Determination if the work is protected by copyright; • Type of work; • License in place and license needed; • Obtaining the license.
Subject	copyright; copyright licensing
Relation	**http://www.athenaeurope.org/getFile.php?id=335** (ATHENA deliverable D6.1)
	http://www.athenaeurope.org/getFile.php?id=665 (ATHENA deliverable D6.3)

3. Conclusions

The ATHENA project has created many tools for the helping organisations, particularly museums, to provide access to their content through Europeana. They include:

- Advice and best practice guide (documents and a step-by-step);
- Metadata standard (LIDO);
- Wiki (terminology);
- Mapping tool and OAI-PMH repository.

All these are online and, in case of the guides, are also available in printed form.

This body of work exemplifies two principles that the ATHENA network wishes to promote:

Continuity and Reuse

In the vernacular:

- "*Let's not reinvent the wheel*" and
- "*If it ain't broke don't fix it!*"

Or to quote Isaac Newton[12]:

- "*If I have seen a little further it is by standing on the shoulders of Giants.*"

To give three examples where the ATHENA project has implemented these principles:

- **Standards advice**

  All the technical standards recommended by the ATHENA project are existing standards. This is not surprising, as there was no need to create new ones. However the purpose of the project was to make the information more accessible to a general cultural heritage audience.

  The project also recommended an existing set of guidelines (those of the *Minerva* project) as an existing solution which meets the needs of the community.

- ***LIDO* metadata harvesting schema**

  LIDO was created by the ATHENA project as a new standard. This was done because it was felt that there was a need for such a standard in the area of giving rich museum metadata to portals, like Europeana.

  Though new, LIDO was built on the firm foundations of existing cultural heritage standards (*CDWA*, *museumdat*, *CIDOC-CRM*, and *SPECTRUM*), and uses the XML technical standard.

  Also LIDO now has an independent 'life' after the ATHENA project. Its home is the *Data Harvesting and Interchange Working Group* of CIDOC, the International Committee for Documentation of ICOM. Here existing ATHENA partners, together with others outside the project, and outside Europe, are working to maintain it and promote its use.

[12] In a letter to Robert Hooke, written on 5th February 1676.

- ***ATHENA Project Ingestion Server***

  The System was developed using Open Source technologies, which re-use existing code. This means that it easy to deploy and adapt the System for use elsewhere.

Both LIDO and the ATHENA System are being used elsewhere:

- In other current ICT PSP projects;
- In the new *Linked Heritage* project;
- In forthcoming project proposals;
- Europeana, and other portals, are considering their use.

Annex 1: A guide to providing content to Europeana

Introduction

The purpose of this document is to give a general introduction to the processes that take place when an organisation is actively engaged in providing content to Europeana. It was written in the context of the ATHENA project, but much of its content to relevant to those not using the *ATHENA Project Ingestion Server* (ATHENA System). In addition it contains handy hints in the form of:

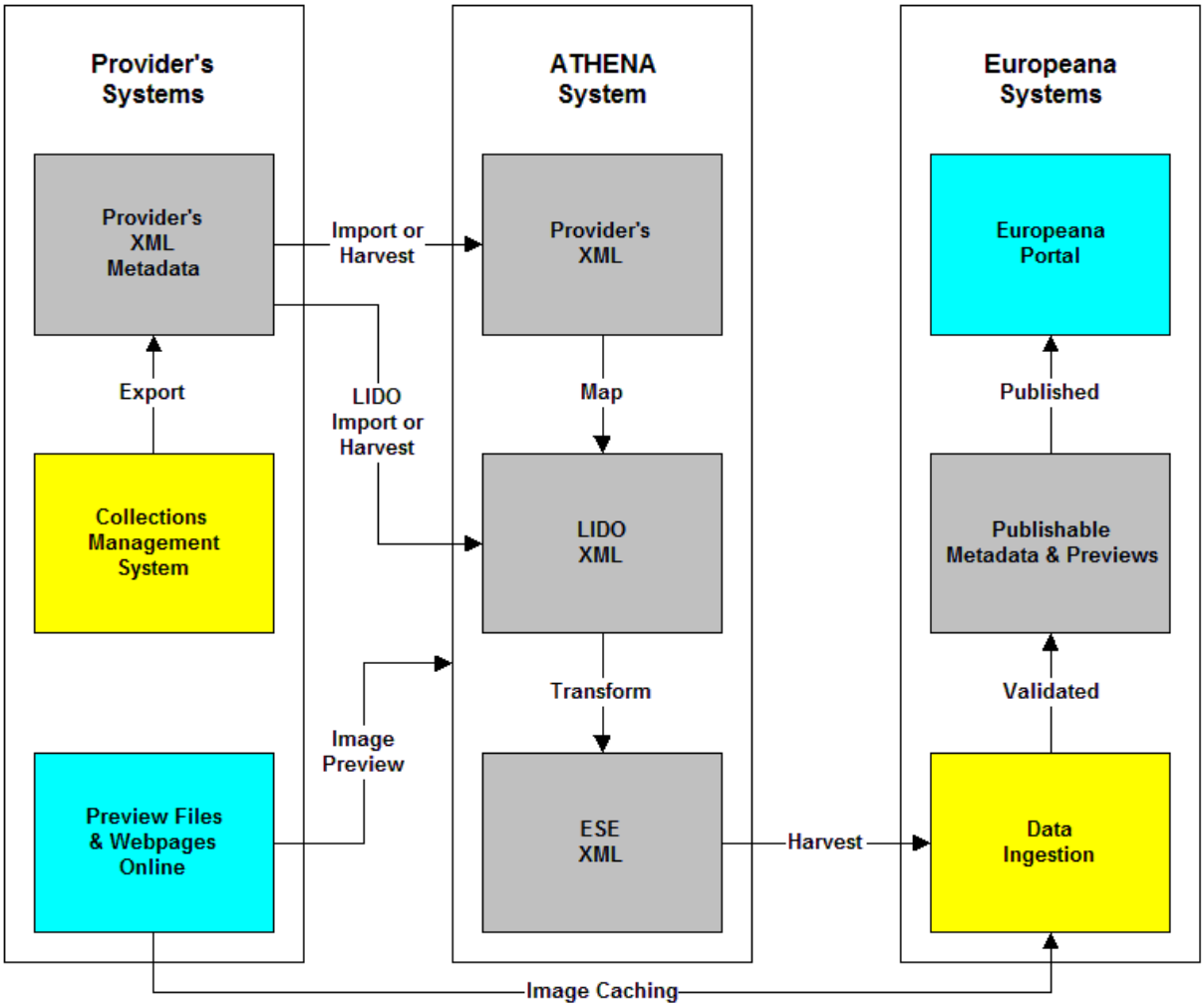
- ***Top tips*** – Important pieces of advice which will lead to the best results;
- ***Bewares*** – Things to look out for and avoid, if possible;
- ***Legal alerts*** – The legal aspects of providing material to Europeana.

These hints are indicated in the text by appearing in boxes with coloured headers.

Note that the document was written in April 2011 and reflects the situation regarding Europeana and the ATHENA System at that date. It will, however, look ahead to some things that were proposed at the time of writing.

Overview of the process

The diagram below is a simplified representation of the provision of material to Europeana using the ATHENA System:



The provision process is divided into three metadata handling 'environments' representing the information systems in three places:

1. ***Provider's Systems***;
2. ***ATHENA System***;
3. ***Europeana's Systems***.

These environments are connected by data transmission 'pipelines' with a flow:

Provider's Systems ⇨ ATHENA System ⇨ Europeana's Systems

This situation is the same for any other aggregator and therefore the ATHENA System can be replaced by:

Provider's Systems ⇨ Aggregator's System(s) ⇨ Europeana's Systems

In the future it is envisaged that the flow of data will be two way. Provider systems will be able to access, and import, enriched metadata into their own systems.

The 'pipe' in this model is the Internet, which is governed by its own set of standards and protocols. These are non-specific for the Europeana ecosystem.

Continuing the analogy, the preferred 'pump' is the *Open Archives Initiative Protocol for Metadata Harvestin*g (OAI-PMH)[13]. This is mechanism that Europeana, and other aggregators, uses to obtain the XML[14] metadata that it needs. However data can 'flow' using other mechanisms such as HTTP[15] upload and FTP[16].

What 'flows' in the pipes is information structured into XML. The XML that flows should be standardised syntactically and, ideally, semantically interoperable.

Europeana enforces syntactic standardisation by requiring providers (actually their aggregators) to deliver XML that conform to its standards, ESE and soon EDM (Europeana Data Model), for XML. These have some mandatory elements. There is also some potential semantic standardisation in the XML elements, particularly where working URIs, usually URLs, are required. Also the **`date`** element and the **`lang`** (language) attribute should conform to standard codes for that kind of data.

Provider's Systems

There is no standard provider's system. Rather there are a set of different solutions to the tasks of collections management and collections access in any organisation. Parts of it may not be computerised at all, or only partly computerised. There may be a set of, perhaps imperfectly connected, systems which are used to manage particular tasks.

It is rare that there is one integrated system for all the processes involved, and even rarer that an organisation's system has the functionality that allows for the automatic harvesting of metadata by OAI-PMH. The diagram above shows 'typical' situations.

TOP TIP – INTEGRATED SYSTEMS
When specifying a new collections management system an organisation should aim for the greatest amount of integration as possible. This will make the work of an organisation in managing and giving access to its collection simpler. If OAI-PMH is also implemented it will make the process of providing material to Europeana (via an aggregator) even more straightforward. Obviously integration will be limited by the cost of such a system.

It is often the case that online access to an organisation's collections is separated from its core collections management system. The former normally contains two 'elements' that are needed for successful provision of material to Europeana:

---

[13] See **McKenna & De Loof**. (2009) *Digitisation: Standards Landscape for European Museums, Archives, Libraries*, (p83).

Download at: **http://www.athenaeurope.org/getFile.php?id=435**

[14] *Ibid*, p51.

[15] *Ibid*, p87.

[16] *Ibid*, p88.

- ***Full information web page*** – This will enable the user to see the material in a user friendly way and go on to explore the rest of the provider's collections. Access to this page is via a URL link. In the ESE the link is contained in the **`isShownAt`** element;
- ***Media file*** – Typically this is a digital image, but can be another media type, which acts as a surrogate for a physical thing. However some things are 'born-digital' and therefore are not surrogates, except in the sense that they are copies of a preserved, and perhaps higher quality digital version. Access to this file is via a URL. In the ESE the link is contained in the **`isShownBy`** element.

The provider has 3 options when providing links, **one of which must be chosen**:

Option	Link(s) provided	Notes
1	**`isShownAt`**	However the **web page being referenced must have a media file accessible in it**. For an image file this is usually a thumbnail which links to a larger image. Europeana cannot access media files on web pages. Therefore a provider wants an image to appear in the Europeana portal it must either chose Option 3 or provide a link to a thumbnail image. In ESE this link is contained in the **`object`** element. The media file referenced in this element is harvested by Europeana, and after possible resizing, is 'cached' in the Europeana systems. The ATHENA System also harvests images from this element for previews of Europeana in the system.
2	**`isShownBy`**	Although this is allowed it is not recommended because a user will not have access to a provider's content except through Europeana. If there is no **`object`** element supplied then Europeana uses this element instead for it 'previews'.
Option	**Link(s) provided**	**Notes**
3	**`isShownAt`** & **`isShownBy`**	An **`object`** element is also an option.

BEWARE – MULTIPLE IMAGES
Quite often an object can only be represented by many images. In Europeana only one image is allowed for each item (i.e. only one **`isShowBy`** (and / or **`object`**) link). Therefore if a provider has multiple images they should choose only one for provision to Europeana. Providers should also consider combining multiple images in a single image. However they

should remember the impact of this on the user of Europeana.

BEWARE – NO THUMBNAILS FROM DJVU FILES

The file format DjVu has been used as an alternative to PDF by some organisations. Europeana can create a thumbnail from a PDF file but cannot do so from a DjVu file. Therefore a provider must either: have a PDF version of the content available for Europeana to create a thumbnail from; have another image file which can be used by Europeana; or not have an image appear on the Europeana portal.

BEWARE – BROKEN LINKS

The provider's online collections access system, and in particular the URLs referenced in metadata, must be available and accessible by Europeana:

- ***At the time of ingestion*** – If links are broken Europeana will reject the records and possibly all the set of records.
- ***While the provider's records are in the Europeana portal*** – If links become broken, as soon as this is detected by Europeana they will as a provider to investigate. Ideally the provider will either fix the problem, or ask Europeana to re-harvest the metadata with the new links. Finally Europeana reserves the right to remove offending records from the portal.

The usual reason for links becoming broken is a change of the system being used to provide access which in turn leads to a change in URLs. Therefore those responsible for maintaining links to Europeana should be made aware, though local communications channels, of changes that are likely to affect this area. Another way to avoid broken links is the use of persistent identifiers (PIDs), see below.

TOP TIP – PERSISTENT IDENTIFIERS (PIDS)

It is highly recommend that providers establish and maintain persistent identifiers in their collection management and access systems for all of their physical and digital objects. This will mean that even if the computer system is changed the links to web pages and media files will not become broken. See the tools discussed above (see *Section 2.8* above for more information).

The next issue to look at is the creation of the XML that will be ingested by Europeana via an aggregator. The simplest solution would be that the XML is directly harvestable from a provider's systems over the Internet, and that it is in the correct form, in this case LIDO. However the diagram above shows the more common situation.

In this situation an XML file is generated by the provider's collections management system which has to be gotten in the ATHENA System.

ATHENA System

In simple terms the ATHENA system is a 'dark aggregator' for museums wanting to provide access to their collections through Europeana. As an aggregator it offers services to its users (the ATHENA partners) which are similar to other aggregators. It is 'dark' because it does not have a public facing service offering access to the material it holds.

The services it offers are:

- ***A relationship between the aggregator and its users***.
  This relationship is unusual because it is set in the context of a time limited, EC-funded, project[17]. However it is similar to other aggregators in that users can register and make use of the services, including being able to manage their 'profiles' and the profiles of the human users that the partner is responsible for.
- ***The ability to import data into the aggregation.***
  The user of the System can upload (or have harvested) its metadata to the aggregation in a number ways. The ATHENA System is quite unusual because it also offers services that allow its users to import non-standard XML to the aggregation. Once there users can manipulate it into the standard format, in this case LIDO. This is then transformed into ESE which can be harvested by Europeana. In addition the system allows the user to see what it will look like in Europeana.

  Users can also directly import LIDO metadata directly and so miss out the data manipulation stage. This option is more like the situation with most aggregators where a standard format is required.
- ***Make metadata available to Europeana.***
  The ATHENA system does by running an OAI-PMH repository, which supports the feature (called 'sets') that allows Europeana to harvest metadata from individual providers, and also parts of the collections of providers.

BEWARE – PUBLICATION IS NOT INSTANT OR AUTOMATIC
When ATHENA partners are satisfied with their mapping of data to LIDO, and with the preview of how it will appear in Europeana, they click on a button which says "publish to Europeana". This is slightly misleading because what it really means is "This set of data is ready to be harvested by Europeana". It is the 'green light' for a series of processes that are not automatic or instantaneous. Once the 'button is pressed' at ATHENA the following happens: 1. The OAI-PMH repository is updated. Once complete then; 2. The Europeana Ingestion Team is e-mailed that a set of metadata is ready to be harvested. Both these processes take time and are only the first stages of publication in the Europeana Portal. It should be noted that any technical difficulties, including those beyond the control of the ATHENA System (e.g. Internet problems), can cause unforeseen delays. (See below for a similar '*BEWARE*' once the data gets to Europeana).

[17] It should be noted that the ATHENA System does have a life after the ATHENA project ends. This is because it will continue to support its users, especially within the context of the Linked Heritage project which began in April 2011 and lasts 30 months.

Europeana's Systems

The ingestion of metadata by Europeana is not automatic – it is 'IT assisted'. This means that although computers, and other IT equipment, are used human beings control their operation. The quality checking of the metadata is one of the key aspects that are human led.

Once a set of metadata records in the ATHENA System are ready for harvesting then the Europeana Ingestion Team (EIT) is informed. The EIT carries out ingestion for all the Europeana Group projects and other aggregators. Therefore the harvesting of an ATHENA set of metadata will have to be scheduled and may not take place immediately after the EIT is informed.

Once the EIT is ready it will harvest the metadata set, as ESE records, from the ATHENA System's OAI-PMH repository. This is relatively fast but can be delayed due to Internet problems. Once the set is harvested then it is checked for compliance to ESE. This should not be a problem because the ATHENA System enforces compliance. However it is possible that some unexpected errors have been introduced. Once the metadata has passed the first test it is checked for conformance to the data content requirements of ESE:

BEWARE – DELAYS DUE TO FAULTY LINKS

The biggest sources of delay in ingestion are related to links, and include issues such as:

- Broken links – URLs should be checked in all records before submission to Europeana.
- Links to the same URL for many items of content – usually the provider does not have a way to generate URLs for individual pieces of content. The provider must find a solution to this before attempting to submit content for access through Europeana.
- Links to content with 'no image' images – Content accessible via must have a digital object available. The usual reason for this situation in that the image has not been copyright cleared. Best advice is for this kind of image not to be present on the provider's system. The information should be presented as text.

With all these issues the EIT will not proceed with publication until they are resolved.

Other errors are due to the wrong data values appearing in elements with a fixed set of allowed values, e.g. **`europeana:type`**. Finally some data just appears in the wrong element. This is harder to detect in the multilingual environment of Europeana.

Once the metadata has been checked the files, accessible via URLs in the **`object`** or **`isShownBy`** element, can be retrieved from the provider's website or Internet accessible system. Once retrieved these files are processed and cached to act as the previews for the Europeana portal. This process can take a considerable amount of time. Factors affecting the speed of preview caching are: the size of the files being retrieved; and the speed of the internet connection between Europeana and the provider.

The final processes involved the creation of indexes and final publication onto the Europeana portal. These are not carried out by the EIT, but take place at regular intervals. The intervals are usually about every two weeks.

BEWARE – APPEARING IN THE EUROPEANA PORTAL TAKES TIME
As can be seen the description above the processes for publishing metadata are time-consuming, and delays can occur at all stages. They occur especially because of the quality checking that has to take place. It is much better if the provider is sure of the quality of the ESE metadata it is proposing to submit to Europeana. However they can also occur due to technical difficulties beyond the control of the Europeana Ingestion Team. Providers should be patient, and contact their aggregator for an update rather than going directly to Europeana.

We now look at each of the three data handling 'environments' in turn in order to help those who want to provide access to their content. This will be done by exploring the key issues, highlighting the pitfalls, and suggesting some solutions.

ESE and Europeana portal

It may seem strange to start at the end of the process, but providers need to be aware of how Europeana displays the ESE[18] (European Semantic Elements) data it harvests. This is because providers will have to make decisions about what metadata they give to Europeana.

What is in ESE

The current ESE (version 3.4)[19] is based on the Dublin Core (DC) metadata element set, and is made up of:

- 15 original DC elements[20];
- 21 elements that are a subset of the DC terms[21], which are refinements of DC;
- 13 elements which were created to meet Europeana's needs.

Here is a summary of the DC and DC terms elements:

DC element	DC terms refinement	Notes
`title`	`alternative`	Can be a very short description.
`creator`		This is the creator of the original item, not the surrogate.
`subject`		These should be descriptive terms from a controlled vocabulary.
`description`	`tableOfContents`	Ideally this should not be a description of

[18] The acronym ESE is usually pronounced like the English word 'easy'.

[19] See: http://www.europeana-libraries.eu/c/document_library/get_file?uuid=77376831-67cf-4cff-a7a2-7718388eec1d&groupId=10128

[20] See: http://dublincore.org/documents/dces/

[21] See: http://dublincore.org/documents/dcmi-terms/

		what can be seen.
publisher		Not the holding institution. Should only be used for publishers of texts.
contributor		Others involved with the physical object, including: engravers of prints, users of objects, and illustrators of books.
date	**created; issued**	Ideally should be standardised. Usually used for creation date, but can be used for any significant dates in an objects 'life', e.g. use.
type		For museum objects this would hold terms like: photograph, painting, sculpture, vase, and coin.
format	**extent; medium**	**medium** should be used for the material(s) the physical object is made of. **extent** should be used for the dimensions of physical objects.

DC element	DC terms refinement	Notes
identifier		Can be the identifier of the physical object, e.g. its acquisition number.
source		Some other thing that the item is derived from, e.g. the painting that a print is derived from. Not the holding institution.
language		Should be used only for the language(s) of text. Not for the language of the metadata.
relation	**isVersionOf; hasVersion; isReplacedBy; replaces; isRequiredBy; requires; isPartOf; hasPart; isReferencedBy; references; isFormatOf; hasFormat; conformsTo**	Mostly used for texts.
coverage	**spatial; temporal**	Not for creation date.
rights		Statement here should not be the same as in **europeana:rights**.
	provenance	A statement of any changes in ownership and

		custody of the item being described since its creation. Not for the holding institution.

BEWARE – LANGUAGE CONFUSION
It is easy to be confused between the language of a text (e.g. a book) and the language the metadata describing an item. However there is a simple rule in the context of Europeana: • Language(s) of texts appears in **`language`** elements; • Language of metadata appears in the **`lang`** attribute of elements. So in the example of a text written in French with a metadata record in English one should see XML like: **`<language>fr</language>`** and **`<description lang="en">This letter describes the meeting between two Frenchmen in London.</description>`** and (perhaps in the same record) **`<description lang="fr">Cette lettre décrit la rencontre entre deux Français à Londres.</description>`**

The **`europeana`** elements of ESE hold data that are used to:

1. Provide links to: the digital object surrogate; the web page describing the object; and a preview (usually a thumbnail) of the digital object surrogate;
2. Display the names of the aggregator, and the organisation holding the original object;
3. Display rights information about the digital object which will allow users to be clear about what they can do with the digital objects accessible through Europeana;
4. Create search facets on the Europeana portal. Most of these are provided by Europeana internally. However some must be given by the provider.
5. Provide terms that are not displayed in the Europeana portal, but can be searched on.
6. Store comments from registered users of the portal
7. Hold the unique identifier for an object in the Europeana system.

Here is a summary of the Europeana elements arranged by these types:

`europeana` element	**Notes**
Type 1 – Links	
`isShownBy`	A URL link to the digital object surrogate.
`isShownAt`	A URL link to a web page describing the object.
`object`	A URL link to a preview of the digital object.
Type 2 – Aggregator and holding institution	
`dataProvider`	The name of the organisation supplying records to the aggregator, i.e. the holding institution of the original physical object.
`provider`	The name of the aggregator.

Type 3 – Rights information

rights	The provider must give one of the terms given in *Guidelines for the* ***europeana:rights*** *metadata element*[22].

Type 4 – Europeana facets

country	Supplied by Europeana. Should be the country of the institution holding physical object.
language	Supplied by Europeana and derived from the **dc:language** element.
type	The provider must give one of (using upper case letters): **TEXT**; **IMAGE**; **SOUND**; **VIDEO**. Even if a provider is giving access to an image of a text (e.g. a book) the **type** is **TEXT** not **IMAGE**. At the time of writing there are no other allowable terms, therefore for other types of digital object (e.g. 3D models) providers should choose the most appropriate.
year	The provider may give multiple values. If not provided Europeana will derive values from data supplied in the **date** element.

europeana element	**Notes**

Type 5 – Unseen search terms

unstored	These are given by the provider, and should be derived from a standard terminology.

Type 6 – Comments by registered Europeana users

userTag	Holds comments from users of the Europeana portal.

Type 7 – Unique identifier in Europeana

uri	Europeana creates this.

BEWARE – TAG ABUSE

Providers must resist the temptation to put data into an ESE element either because it is has a name similar to a provider's system or it seems to be 'spare' and not used for anything else.

This is called 'tag abuse' and may also occur within a provider's own system.

Best practise is for providers to conform to the ESE standard as published by Europeana[23], and if in any doubt they should contact their aggregator.

BEWARE – MULTIPLE ELEMENTS AND ORDER

[22] See: http://www.europeana-libraries.eu/c/document_library/get_file?uuid=06e63d96-0358-4be8-9422-d63df3218510&groupId=10602

[23] *Metadata Mapping & Normalisation Guidelines for the Europeana Semantic Elements.*

See: http://www.europeana-libraries.eu/c/document_library/get_file?uuid=b3cfcf47-da0a-4c6b-b1d7-9b08e162643e&groupId=10128

In keeping with standard Dublin Core practise many of the elements in ESE are repeatable. However this does not mean that the order of these repeating elements will be preserved in any application using the XML data. This is the case for the Europeana portal. The next section describes how the Europeana portal uses ESE data.

How the Europeana portal displays ESE

The Europeana portal displays the ESE XML data that it harvests to create a portal which gives the citizens of Europe, and beyond, access to their cultural and scientific heritage. In order to do this it presents the data in it holds in a number of 'displays'.

Below we look at the two main displays. In particular we highlight what happens when some elements in ESE are repeated, and what happens when some elements are missing[24]:

***Brief display*:**

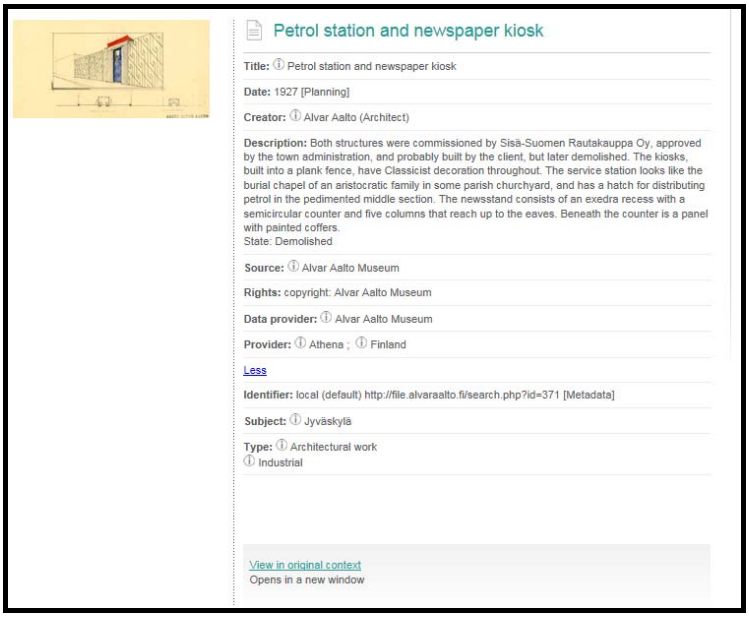
Used in the search results and the header of the ***Full display***.

Display heading (in order)	ESE element displayed (in preferential order[25]) with notes
Title	1. **`dc:title`** – first occurrence (others not displayed); 2. **`dcterms:alternative`** – first occurrence (others not displayed); 3. **`dc:description`** – first occurrence (others not displayed)
Creator	1. **`dc:creator`** – first occurrence (others not displayed); 2. **`dc:contributor`** – first occurrence (others not displayed)
Date	1. **`europeana:date`** – first occurrence (no options)
Data Provider	1. **`europeana:dataProvider`** – (no options)
Provider	1. **`europeana:provider`** – (no options)

[24] The authors are grateful to the Europeana Ingestion Team for their help in creating this section. Any mistakes are the authors' alone.

[25] This means that Option 1 is implemented first if possible. If Option 1 is not possible then Option 2 implemented, and so on.

***Full display*:**



The default display is with labels. The labels are displayed in the interface user-selected language. The display is in two blocks: always displayed and the 'More' display option.

Always displayed:

Display heading (in order)	ESE element(s) displayed with notes
Title	**`dc:title`** and **`dcterms:alternative`** – All occurrences. Each appears: in alpha-numeric order; and on a new line.
Date	**`dc:date`**, **`dcterms:created`**, **`dcterms:issued`** – All occurrences. Each appears: in alpha-numeric order; in the same line; and separated by '; '.
Creator	**`dc:creator`** – All occurrences. Each appears: in alpha-numeric order; in the same line; and separated by '; '. **`dc:contributor`** – All occurrences. Each appears: in alpha-numeric order; in the same line; and separated by '; '.
Description	**`dc:description`** – All occurrences (which are not used in the title header). Each appears: in alpha-numeric order; and in a new line. Text is limited to 800 characters with the last full word followed by '…'
Language	**`dc:language`** – All occurrences. Each appears: in alpha-numeric order; in the same line; and separated by '; '.
Format	**`dc:format`** – All occurrences. Each appears: in alpha-numeric order; in the same line; and separated by '; '. **`dcterms:extent`** – All occurrences. Each appears: in alpha-numeric order; in the same line; and separated by '; '. **`dcterms:medium`** – All occurrences. Each appears: in alpha-numeric order; in the same line; and separated by '; '.

Display heading (in order)	ESE element(s) displayed with notes
Source	**dc:source** – All occurrences. Each appears: in alpha-numeric order, and starts on a new line.
Rights	**dc:rights** – All occurrences. Each in alpha-numeric order, and separated by '; '.
Data provider	**europeana:dataProvider**
Provider	**europeana:provider**, **europeana:country** Separator is ', '

'More' display option:

Publisher	**dc:publisher** – All occurrences. Each appears: in alpha-numeric order; in the same line; and separated by '; '.
Identifier	**dc:identifier** – All occurrences. Each appears: in alpha-numeric order; in the same line; and separated by '; '.
Provenance	**dcterms:provenance** – All occurrences. Each appears: in alpha-numeric order; in the same line; and separated by '; '.
Subject	**dc:subject** – All occurrences. Each appears: in alpha-numeric order; in the same line; and separated by '; '.
Coverage	**dc:coverage**, **dcterms:spatial** and **dcterms:temporal** – All occurrences. Each appears: in alpha-numeric order; in the same line; and separated by '; '.
Type	**dc:type** – All occurrences. Each appears: in alpha-numeric order; in the same line; and separated by '; '.
Relations	**dc:relation**, **dcterms:isVersionOf**, **dcterms:hasVersion**, **dcterms:isReplacedBy**, **dcterms:replaces**, **dcterms:isRequiredBy**, **dcterms:requires**, **dcterms:isPartOf**, **dcterms:hasPart**, **dcterms:hasFormat**, **dcterms:conformsTo**, **dcterms:isReferencedBy**, **dcterms:references** – All occurrences. Each appears: in alpha-numeric order; and in a new line; and separated by '; '.

BEWARE – DISPLAY IN ALPHA-NUMERIC ORDER

Europeana shows, in the *Full display*, multiple occurrences of the same element in alpha-numeric order, often on the same line, separated by '; '. This also is the case where the same record has multiple language versions for the same element, even if it is marked with an appropriate **lang** attribute.

Therefore providers should test, by using the preview in the ATHENA System, what will happen to their metadata in the Europeana portal. They must then take a decision on if they are happy with this, or take some steps to alleviate the situation (see next two *TOP TIPs*).

TOP TIP – CONCATENATING DATA

The *BEWARE* above highlighted the issue of what happens to repeating elements in the Europeana portal – they appear in alpha-numeric order, often with a separator. Some providers of course can accept this situation. Others, however, may take the view that it will compromise the meaning of their metadata. In this case they have two options:

1. Do not submit that part of their metadata to Europeana;
2. Concatenate their metadata into one ESE element, and mimic the Europeana separators.

Ideally this concatenation should be done within the provider's own systems. However it is also possible to do this within the ATHENA System.

TOP TIP – MULTILINGUAL METADATA

The *BEWARE* above highlighted the issue of what happens to repeating elements in different languages – they appear in alpha-numeric order, often with a separator. This may result in 'strange' records that a provider may find unacceptable. In this case the provider has two options:

1. Limit themselves to providing one monolingual record per object;
2. Provide many monolingual records in different languages in different sets.

Option 2 is especially appropriate where the **`isShownAt`** element is not the same for the different languages.

LEGAL ALERT – PROPOSED EUROPEANA DATA PROVIDER AGREEMENT

The current licence that Europeana has with its providers restricts the use of the metadata it receives to non-commercial use.

Europeana wishes to publish the metadata it has, with enrichments, as Linked Open Data. In order to do this its providers will have to agree to the removal of the non-commercial clause from the agreement. This does not apply to the previews that Europeana gives, only to the metadata.

This agreement is likely to be introduced in latter part of 2011.

The provider's metadata handling environment

We next consider the situation within the provider's metadata handling environment. This is the most important area for the set of processes that need to work successfully, in order to give metadata to Europeana. Some issues for a potential provider to consider looking at are:

The provider's systems

As was discussed above providers' collections management and access systems can vary enormously. Some can be highly integrated, with customisable metadata in different XML formats available at the 'touch of a button', or even harvestable via OAI-PMH. The worse-case scenario might be a limited system, with very few parts of it computerised, but with an online presence

provided by someone else. Most potential providers are somewhere between the two extremes. Therefore:

TOP TIP – KNOW YOUR SYSTEMS

Those who are working to provide access to their collections through Europeana need to know their own systems in order to carry out the task effectively and efficiently. They need to be aware of:

- ***The systems' limitations*** – What it cannot do;
- ***The data not stored in the system*** – This might be implicit information or knowledge is somebody's head;
- ***The metadata in the system that should not be given to Europeana*** – Usually security sensitive or personal;
- ***The tools available to check the quality and consistency of metadata*** – See below.

BEWARE – NO DIGITAL CONTENT AVAILABLE ONLINE

Providers can only give metadata records to Europeana where they can give access to a digital media file over the Internet, and usually over the Web. Records that do not meet this requirement will be rejected by Europeana, usually at ingestion time. When the prototype Europeana portal was launched there was some material like this, but this in now being eliminated.

The ATHENA project does allow partners to upload records where there is no digital media file. This is because it can be argued that records without digital media files, but with good metadata, are useful for a general user to know that such items exist. Users of the ATHENA system should be careful not carry out LIDO mappings which create 'false' URLs to non-existent digital media files, or web pages without those files.

TOP TIP – KEEP IT SIMPLE

Remember that you need not give all you metadata to Europeana. Reasons for not giving certain metadata are, that it is:

- Aimed at a specialist audience which Europeana is not focussed on;
- Not possible to represent well in the Europeana portal (its meaning is lost);
- Difficult, or impossible, to export from the provider's system;
- Not available for publication as Linked Open Data.

Data quality and consistency

There is assumption that the data in the provider's systems is of good quality and consistent. Unfortunately even with the best run systems it is possible for things to go awry. Unless data entry is very tightly controlled and monitored then a number of issues tend to cause errors. These include metadata:

- With spelling mistakes;
- Not conforming to internal or external rules (e.g. 'XXth century' rather than '20th century');
- In the wrong field, due to data entry error;
- In the wrong field deliberately ('tag abuse');
- With inconsistencies due to changes in practise over time and unmanaged data entry staff.

Such issues will need to be fixed before the metadata is submitted to Europeana. Therefore there will probably need to be at least one phase of data quality checking and updating before it is ready for export.

TOP TIP – CHECK YOUR METADATA WITH THE ATHENA SYSTEM
If an ATHENA partner has difficulty in checking the quality and consistency of data in their own environment then the ATHENA System has a simple way to check the data once it has been imported. For each data element imported it is possible to press the '  ' button to see a listing of values, with frequencies in that element.

Planning, staff training and communications

TOP TIP – PLAN YOUR WORK
It is good practise to have some kind of plan in place before carrying out a project to submit metadata to Europeana. This work might not merit a full implementation of a project management methodology. However like the ATHENA project it would be useful to have: • ***Objectives*** – What the provider is trying to achieve; • ***Tasks*** – How the objectives will be met; • ***Deliverables*** – What will be the outcomes; • ***Milestones*** – Significant points along the way; • ***A visualisation of the plan*** – Perhaps a Gantt chart; • ***Resources*** – Personnel, tools available, time, and money; • ***Success indicators*** – measurable and relevant; • ***Risks*** – What might go wrong, how likely is it, and what will be done to mitigate them.

It is important for an organisation not to neglect the personnel ('human resources') and communication aspects of the project. Those involved, including senior management, should be aware of what is being planned. They should know about the context of the project, their role within it, and the channels of communications that exist. The tools created by the ATHENA project will play a key part in this information system.

Rights to display previews in the Europeana portal

This is a major issue that a provider should address:

LEGAL ALERT – IPRS

Providers **must** give metadata about the rights status of the original content they are giving access to through Europeana.

These legal rights are associated with material in the provider's collections (the objects themselves, photographs and other surrogates for the objects, and the descriptions of objects). These rights include, **but are not limited to copyright**. Providers **must** manage and document the rights associated with their collections, in order to benefit the organisation and to respect the rights of others.

Permission to use copyrighted content, technically called 'works', can be given by a licence. Many organisations are experienced in copyright licensing for their own use, e.g. on their website.

However a provider **must not** infringe the rights of others when it provides content to Europeana. Agreements with Europeana specifically state that providers must guarantee that they are not doing so. Providers must also indemnify Europeana for any legal claims where they are, in fact, infringing rights. This gives Europeana 'clean hands', and is normal practice.

TOP TIP – SOLUTIONS TO IPRS ISSUES ASSOCIATED WITH PREVIEWS

There are two situations where a provider can infringe the rights of others. This concerns the previews that Europeana displays on its portal:

- The provider has a licence for the content to be used on their own website, but it does not cover use by a third party (i.e. Europeana). There are two options:
    - Obtain a licence which allows them to provide the content to Europeana; or
    - Do not provide the content to Europeana. This is allowed so long as the provider can give access to the content on their website. Therefore the preview will not appear in Europeana, but the metadata about it will.
- It is not possible for the provider to obtain a licence. This is the situation of so-called 'orphan works', where the creator of the original content, or the current owner of the copyright, is unknown. Use of the content by a provider themselves is governed by managed 'risk'. If an organisation provides this kind content to Europeana then the risk is the provider's alone. It is not shared with Europeana.

In general terms the provider must have policies and guidelines on rights. It is good practice for these to state the:

- Steps to be taken to research rights associated with works that become (or might become) the responsibility of the organisation;
- Steps to be taken to ensure the protection of the rights of the organisation and the rights of others;
- Response to claims to breach of rights both against and on behalf of the organisation;

Ability to produce XML

One of the major requirements of providing metadata to Europeana is for that metadata to be in XML. Most modern collections management systems can produce 'XML' as an output. However providers should check their 'standard' XML export to see what data it contains, and what data it does not contain. Difficulties that can occur include:

- ***'XML' is not valid*** – See the Beware below for how you can check;
- ***Sensitive data*** – The data may include collections management information that should not be passed on to Europeana (e.g. locations of objects and personal data of donors).

  Providers should check for this and: either produce an XML export without this data; or make sure they do not map this data within the ATHENA System to LIDO;
- ***Missing data*** – Typically this is: information that is not needed for the provider's system, but is needed for Europeana (e.g. the name of the provider and local place names); and no URL links because they are created by the collections access system based on data such as an object number.

  Providers should check for this and: either produce an XML export which includes the needed data; or use the ATHENA System to add the data when mapping to LIDO.

BEWARE – CHECK YOUR 'XML'
Providers should check the file, that they are attempting to upload to the ATHENA System, is actually XML. It is surprising how small errors can creep in during the export process, e.g. in export scripts. Also if a provider is proposing to import LIDO data directly they should check it against the LIDO schema[26]. If they do not have a specialised XML editor, the quickest way to check for validity is to attempt to view the file in the web browser *Internet Explorer*. This should bring up any obvious errors.

TOP TIP – CONVERTING EXCEL DATA TO XML
Although not recommended as a starting point the authors recognise that some providers are unable to produce their metadata in XML form. However they can export the appropriate metadata in an Excel spreadsheet. There is a simple method to convert simple Excel data into XML. The section at the end of this document, *Converting Excel to XML: A simple method*, explains this.

[26] See: **http://www.lido-schema.org/schema/v1.0/lido-v1.0.xsd**

The ATHENA System

The purpose of this section is not to be a user manual for the ATHENA System. The training that the ATHENA project carried out, and filmed is a much better way to understand the System[27]. Of course one of the best ways to learn about any system is just to 'play' with it. What we do here is to highlight some points which will help the user to produce better metadata for ingestion into Europeana.

The ATHENA System was created by the ATHENA partner NTUA using open source technologies. It is a web-based tool with access controlled by usernames and passwords. With the system ATHENA partners can:

Create and manage organisational and user profiles.
Organisations can have sub-organisations. Users can be assigned to any organisation and have different permissions for viewing and editing files.

TOP TIP – CREATE VIRTUAL SUB-ORGANISATIONS
The original use of sub-organisations was envisaged for the situation where a partner was acting as a 'mini-aggregator' for other organisations. However it also an advantage for a single organisation. One consequence of having sub-organisations is that it data associated with them can be harvested separately, as an OAI-PMH 'set'. It is suggested that an organisation creates 'virtual sub-organisations' of itself, based on the different blocks of records it intends upload the ATHENA System. Mappings can be shared between these different sub-organisations, and modified if needed. The advantage of all this is to speed up the Europeana ingestion process. This is because it is possible only harvest those records associated with the sub-organisation (as an OAI-PMH set). Not having sub-organisations means that there is just one OAI-PMH set for an organisation, and therefore **all** the metadata will have to be harvested by Europeana every time there additional data from by the provider. Europeana cannot incrementally harvest the same OAI-PMH set.

Import their metadata in any XML format.

The system checks that the imported metadata is valid XML. (LIDO XML itself can be imported). Import methods are: HTTP upload, FTP (NTUA and remote servers), and OAI-PMH harvesting.

BEWARE – THE TIME TO UPLOAD CAN BE SIGNIFICANT
The use of the Internet to transport data can be very slow, especially when the uploading data is being carried out. Uploading, or harvesting (by OAI-PMH), metadata to the ATHENA System can take a significant amount of time. Providers can decrease that time by putting their files in ZIP archives (i.e. the files can become considerably smaller). However, depending on the size of the file and the number of records, **be prepared to wait!**

[27] See: **http://www.athenaeurope.org/index.php?en/159/training**

Map their XML metadata to LIDO.
This is most simply done by a one-to-one element mapping, using a 'drag and drop' mechanic. However more complex mappings can be carried out like: conditional mapping; concatenation, and assignment of 'constant values' where data is absent from the imported metadata.

Users of the system can see and have access to the XML transformation control documents (XSLTs). A preview of how records will appear in Europeana is also available.

BEWARE – FALSE URLS

One of the major difficulties that arise with the Europeana ingestion process is because providers, using the ATHENA system, accidently create 'false URLs'. This because the URLs are being created by concatenating data from an element in the uploaded file with 'constant data' added by the user of the System. An example might be:

`http://www.provider.org/collection-opac/1234.jpg`

Only the '1234' is data in an element and the rest has been added as two 'constants' in the ATHENA System. This works well when every record has data but it fails when some of the records do not have data in that element. This will lead to a 'false URL':

`http://www.provider.org/collection-opac/.jpg`

The ATHENA System cannot distinguish this situation and therefore it 'publishes' the record as being available to Europeana. Europeana rejects the record, and possibly the whole set of records as being potentially having broken links.

There are two possible solutions:

- The provider separates the set of records into two sets: those with data in the element, and those without the data in the element. In the latter set mapping does not take place.
- The provider uses the ATHENA system to test if there is data in the element, and only does a mapping in a record where there is. This requires knowledge of how XML transformations (XSLTs) work.

TOP TIP – PREVIEW WHAT YOUR DATA LOOKS LIKE

Always check a reasonable sample of records before attempting to 'publish' them. Look at them in the *Europeana Preview*. Check:

- The general appearance of the records – do they look okay?
- The links to your website – do they work? If they do not investigate why.

This checking may lead to further mapping work. If you are having trouble, ask those more experienced in using the System.

Make their transformed data available to Europeana.
When Europeana has been given permission it harvests the data as valid ESE records, using OAI-PMH, and eventually a provider's records will become visible in the Europeana portal.

BEWARE – VISIBILITY IN EUROPEANA IS NOT INSTANT

This has been mentioned above records appearing in Europeana from your organisation will take a significant amount of time.

Converting Excel to XML: A simple method

Although not recommended as a starting point, the authors recognise that some providers are unable to produce their metadata in XML form. However they can export the appropriate metadata in an Excel spreadsheet. Here we describe a simple procedure that a provider can follow to convert data from Excel to XML. Please note that an assumption is being made that the Excel data is very simple, and is not a complex set of sheets which mimic a relational database.

The process we described can also be viewed as a short video (created by the Mike Lively of Northern Kentucky University)[28]. This video shows the process for versions of Excel earlier than *Excel 2007* that the authors have. The process is largely the same, but what follows refers to *Excel 2007*. Users should refer to the video for the differences.

Step 1: Examine the Excel data

Look at the data in the Excel sheet that you want to convert. It should appear in columns, where the first row gives the names of the fields in the database from which data has been exported. Here is a screenshot of some test data:

	A	B	C	D	E	F	G	H	I	J	K
1	field1	field2	field3	field4	field5	field6	field7	field8	field9	field10	
2	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	
3	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	
4	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	
5	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	
6	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	
7	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	
8	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	
9	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	
10	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	
11	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	
12	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	
13	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	
14	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	
15	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	
16	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10	
17	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	
18	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	
19	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	
20	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
21	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	
22	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	
23	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	
24	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	
25	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	
26	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	
27	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10	
28											

If the first row does not contain field names then add some in a similar way to the example above. The field names from databases are likely to be okay for XML, but if they contain any space characters replace these with '-'. The rows of the spreadsheet represent the records in the database.

[28] See: Mike Lively (2008). *Convert Excel Spreadsheet data to XML*.

View at: **http://www.youtube.com/watch?v=9bat12gH3Qs**

Step 2: Creating an XML schema for importing to Excel

The next step is to create an XML schema which contains data from two rows (records) of the spreadsheet. This can be done in *Notepad*. It should look like this:

```
<?xml version="1.0"?>
<wrapper>
  <record>
    <field1>1A</field1>
    <field2>2A</field2>
    <field3>3A</field3>
    <field4>4A</field4>
    <field5>5A</field5>
    <field6>6A</field6>
    <field7>7A</field7>
    <field8>8A</field8>
    <field9>9A</field9>
    <field10>10A</field10>
  </record>
  <record>
    <field1>1B</field1>
    <field2>2B</field2>
    <field3>3B</field3>
    <field4>4B</field4>
    <field5>5B</field5>
    <field6>6B</field6>
    <field7>7B</field7>
    <field8>8B</field8>
    <field9>9B</field9>
    <field10>10B</field10>
 </record>
</wrapper>
```

Note that the first line is a ‘comment’ which says that this is an XML file.

The rest of the file holds the data. It has:

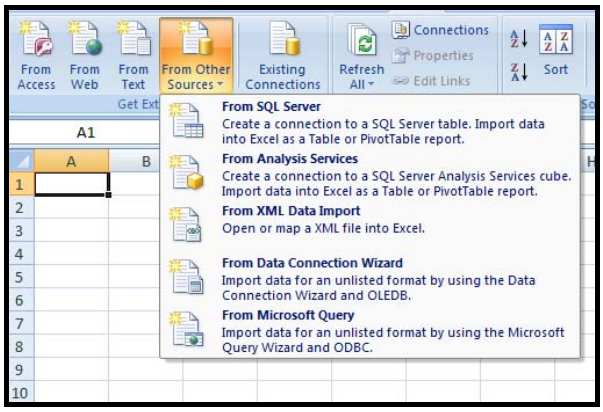
1. A **`wrapper`** element which surrounds the whole document;
2. Each record is surrounded by a **`record`** element. There must be two records in the file;
3. A series of elements (**`field1`**, **`field2`** .... **`field10`**) which hold the data.

The space characters in front of elements are not needed, and are just here to make reading easier.

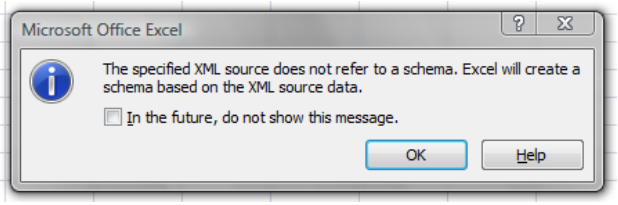
The file should not be saved as ‘**`filename.txt`**’, but as **‘`filename.xml`’**, with ‘Encoding’: **`Unicode`**. It should be checked to see if it is valid XML. This can be easily done by opening it in *Internet Explorer*. Any errors should be obvious.

Step 3: Importing the XML schema into Excel

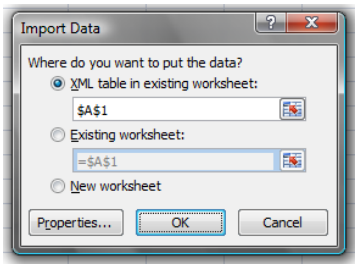
Go back to Excel and click on the tab for an empty worksheet (usually called: ‘Sheet2’). Doing this should highlight cell A1. Click on the ‘Data’ tab and then on ‘From Other Sources’ which then shows this:



Choose ‘From XML Data Import’. This allows you to import the XML you have saved and checked. When you import you may get a message like this:



Click on the ‘OK’ button. This may bring up:



Click on the ‘OK’ button. Doing this will import the schema, and will result in:

	A	B	C	D	E	F	G	H	I	J	K
1	field1	field2	field3	field4	field5	field6	field7	field8	field9	field10	
2	1A	2A	3A	4A	5A	6A	7A	8A	9A	10A	
3	1B	2B	3B	4B	5B	6B	7B	8B	9B	10B	
4											
5											

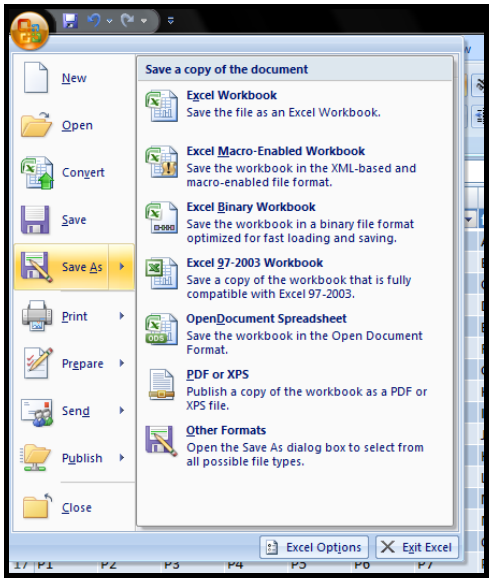
Step 4: Copying the rest of the data

Go back to first Excel worksheet with all the data in it. Highlight this data, but not the field names, and ‘Copy’ it. Go to second worksheet (like above). Click on cell A2 and ‘Paste’. This should result in:

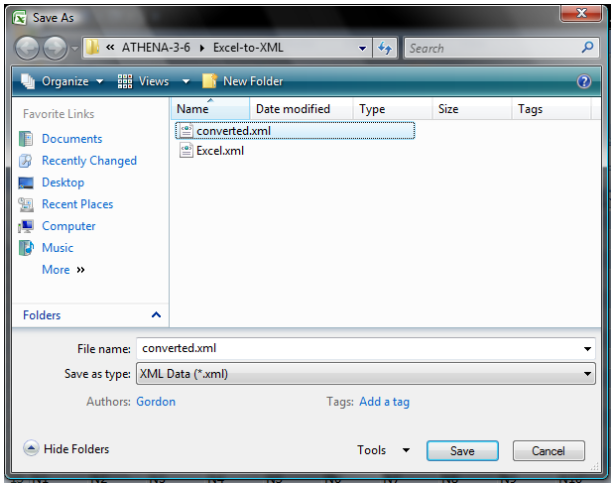
	A	B	C	D	E	F	G	H	I	J
1	field1	field2	field3	field4	field5	field6	field7	field8	field9	field10
2	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
3	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
4	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
5	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
6	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
7	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
8	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
9	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10
10	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
11	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10
12	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
13	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
14	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
15	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10
16	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
17	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
18	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
19	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
20	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
21	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
22	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
23	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
24	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
25	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
26	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
27	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
28										

Step 5: Saving the data as XML

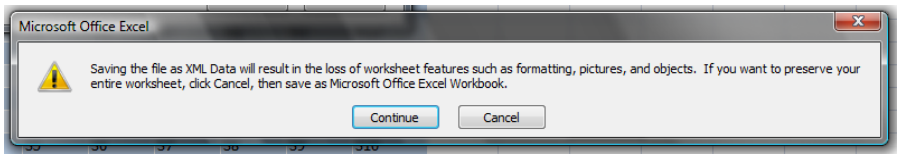
Click on the ‘Office Button’ (top left), and choose ‘Save As’ to see:



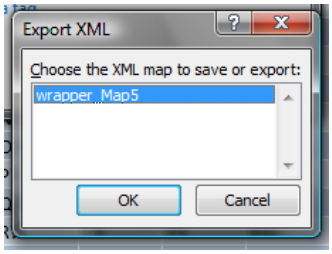
Choose ‘Other formats’. This will bring up a dialogue box which will allow you to choose ‘Save as type’ ‘XML Data (*.xml)’ with an appropriate filename:



Click on the ‘Save’ button, which will bring up:



Click on the ‘Continue’ button, which may bring up:



Click on the ‘Okay’ button. This will save all the data as an XML file ready for uploading to the ATHENA System.