ECDL 2009 Tutorial

Aggregation and reuse of digital objects' metadata from distributed digital libraries

Prepared by: PSNC Digital Libraries Team (http://dl.psnc.pl/)

Speaker: Marcin Werla (mwerla@man.poznan.pl)

Part 3: The most important aspects of… the metadata provisioning

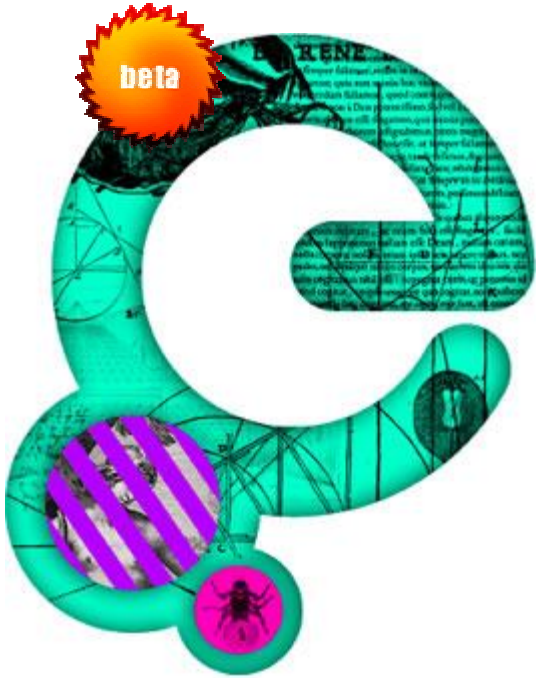
ECDL 2009 Tutorial: **Aggregation and reuse of digital objects' metadata from distributed digital libraries**

Europeana

- Portal which gives access to European cultural heritage
- Information comes from:
  - Museums
  - Archives
  - Libraries
  - Audiovisual collections

Europeana

- First prototype version was enabled on 20.11.2008
- Now Europeana gives access to over 4 millions of digital objects distribute all over the Europe
- Europeana is a „metadata directory"
  - Access to the contents of the digital objects is made on the websites of their origin



europeana
pomyśl o kulturze

Europeana

http://europeana.eu

Europeana

- Main way of financial support for this initiative are projects co-funded by the European Commission
  - Previously under eContent*Plus* programme
  - Now CIP ICT-PSP
    - Theme 2: Digital Libraries
      1. European Digital Library – services
      2. European Digital Library – aggregating digital content in Europeana
      3. European Digital Library – digitising content for Europeana
      4. Open access to scientific information
      5. Use of cultural heritage material for education

europeana
pomyśl o kulturze

Europeana

- Ongoing projects
  - Technical/organizational
    - Europena v1.0
      - Should result in a production-ready version of Europeana
    - Europeana Connect
      - Development of technologies necessary for the Europeana
    - PrestoPRIME
      - Long-term preservation of audiovisual materials

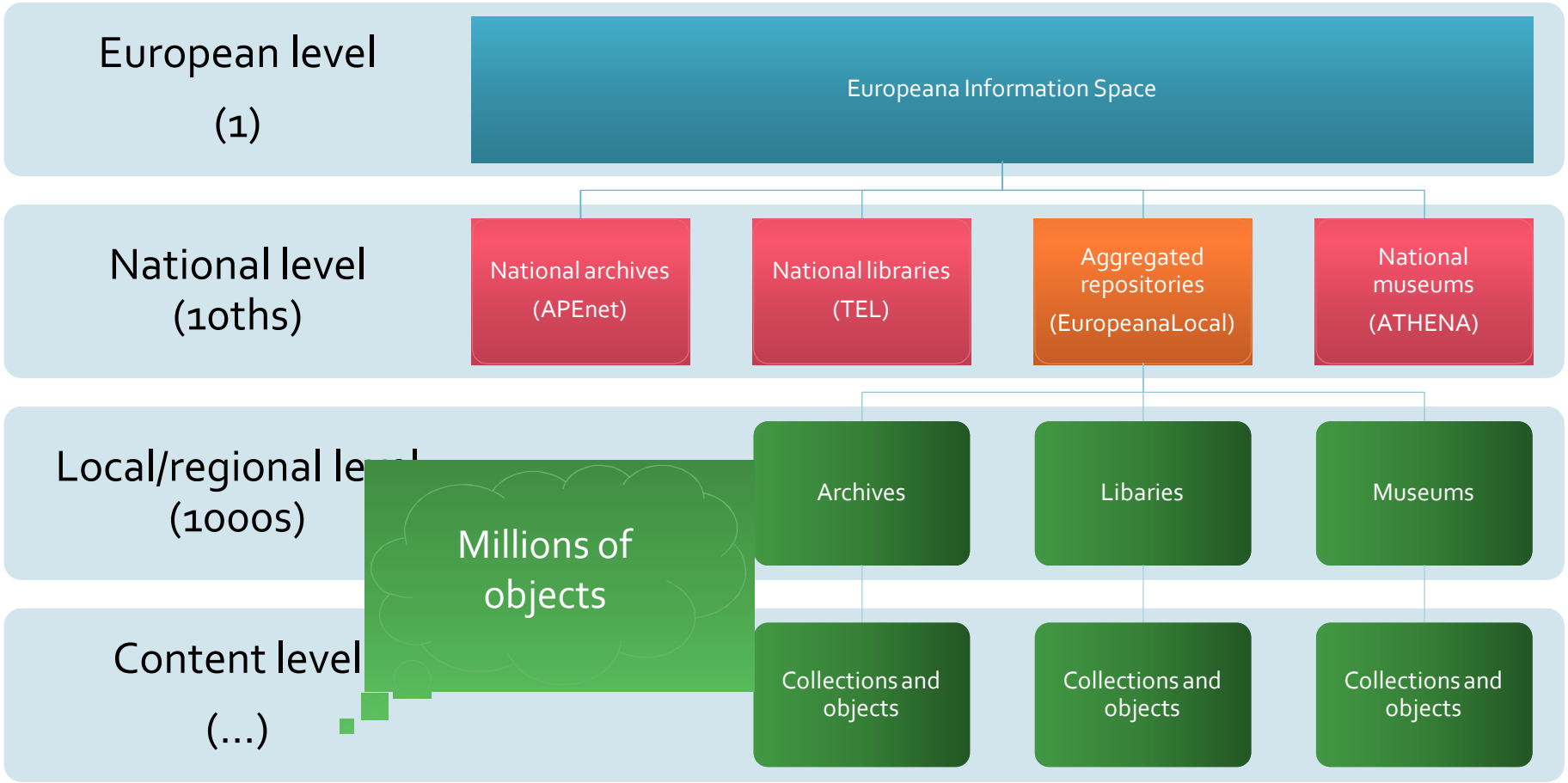
europeana
pomyśl o kulturze

Europeana

- Ongoing projects
  - Content providers
    - APEnet – national archives
    - ATHENA – museums (national level)
    - BHL – Europe – biodiversity heritage library
    - EUscreen – TV materials
    - Europeana Connect – audio materials
    - Europeana Local – materials from local and regional institutions
    - Europeana Travel – travel, tourism, …
    - Judaica Europeana – influence of Jewish culture on European cities
    - EFG – movies/cinema

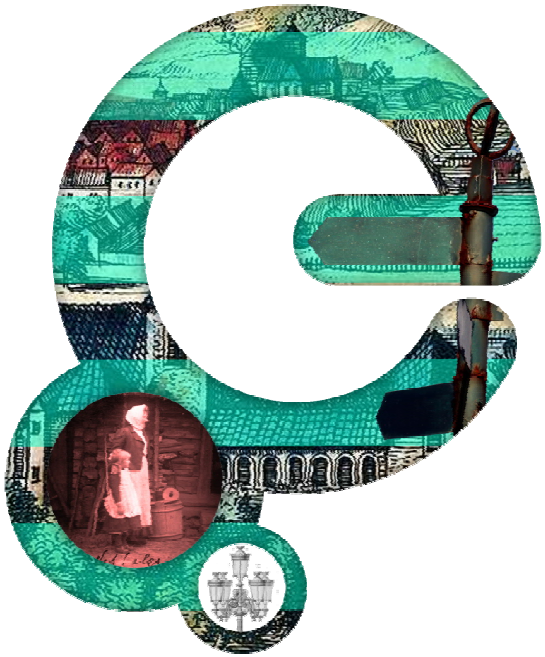
More information at: http://group.europeana.eu/

Local and regional resources in the European information space



EuropeanaLocal

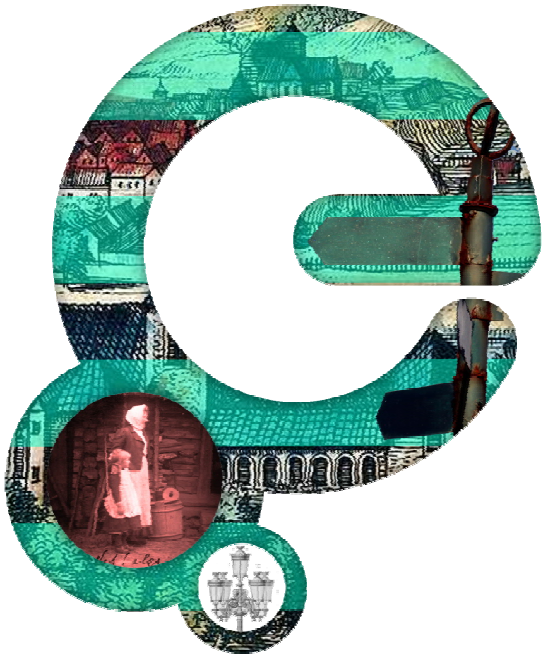
- European project
  - Funded under the eContent*Plus* programme
- Duration – 3 years
  - Since 1 June 2008 to 31 may 2011
- Type of the project
  - Best practice network



europeana
local

Main aims

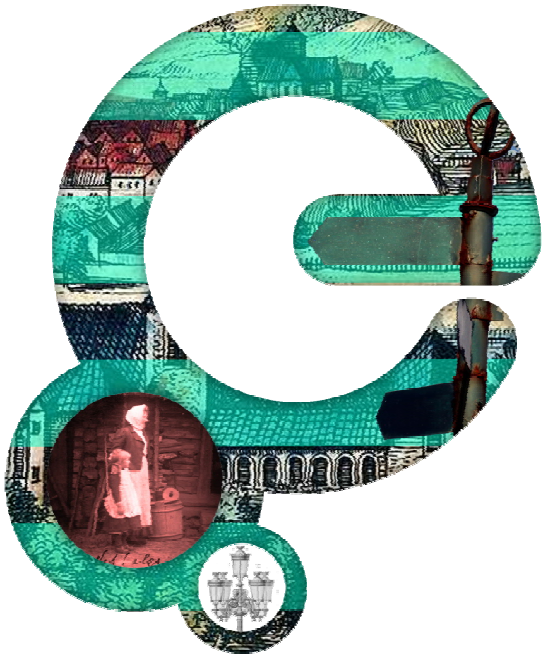
- Improvement of the interoperability of digital content
  - Increase of possibilities for automated use of digital content located in local and regional digital libraries
  - Support for the creation of metadata aggregation services on different levels
- Creation of the network of local repositories being able to communicate with Europeana
  - Support for the creation of infrastructure compatible with Europeana
  - Development of tools and processes which will facilitate the establishment of the cooperation with Europeana in the future
  - In opposition to many other projects, EuropeanaLocal is not going to built its central aggregator



europeana
local

Participants (32)

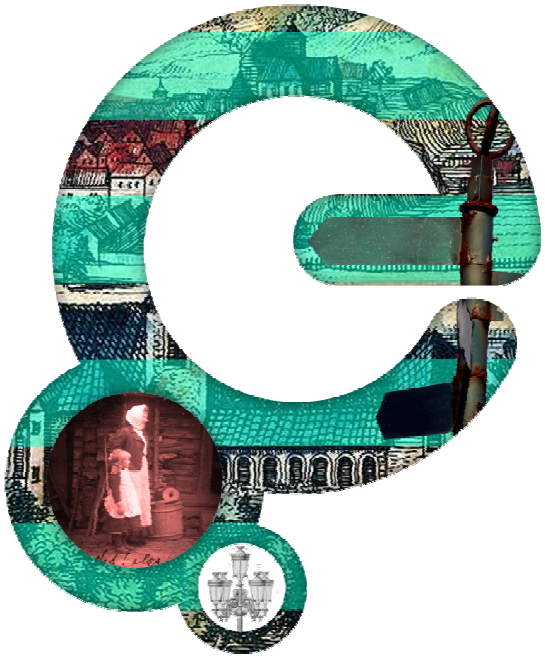
- Project coordinator
  - Sogn og Fjordane County Municipality (NO)
- Management support and scientific cooperation
  - MDR Partners (UK)
- Technical partners
  - EDL Foundation (NL) – main source of standards
  - 3 x technical support (SK, 2 x NO)
- Country coordinators x 27
  - AT, BE, BG, CY, CZ, EE, ES, DK, FI, DE, FR, GR, HU, IE, IT, LV, LT, MT, NL, *NO*, PL, PT, RO, SK, SI, SE, UK



europeana
local

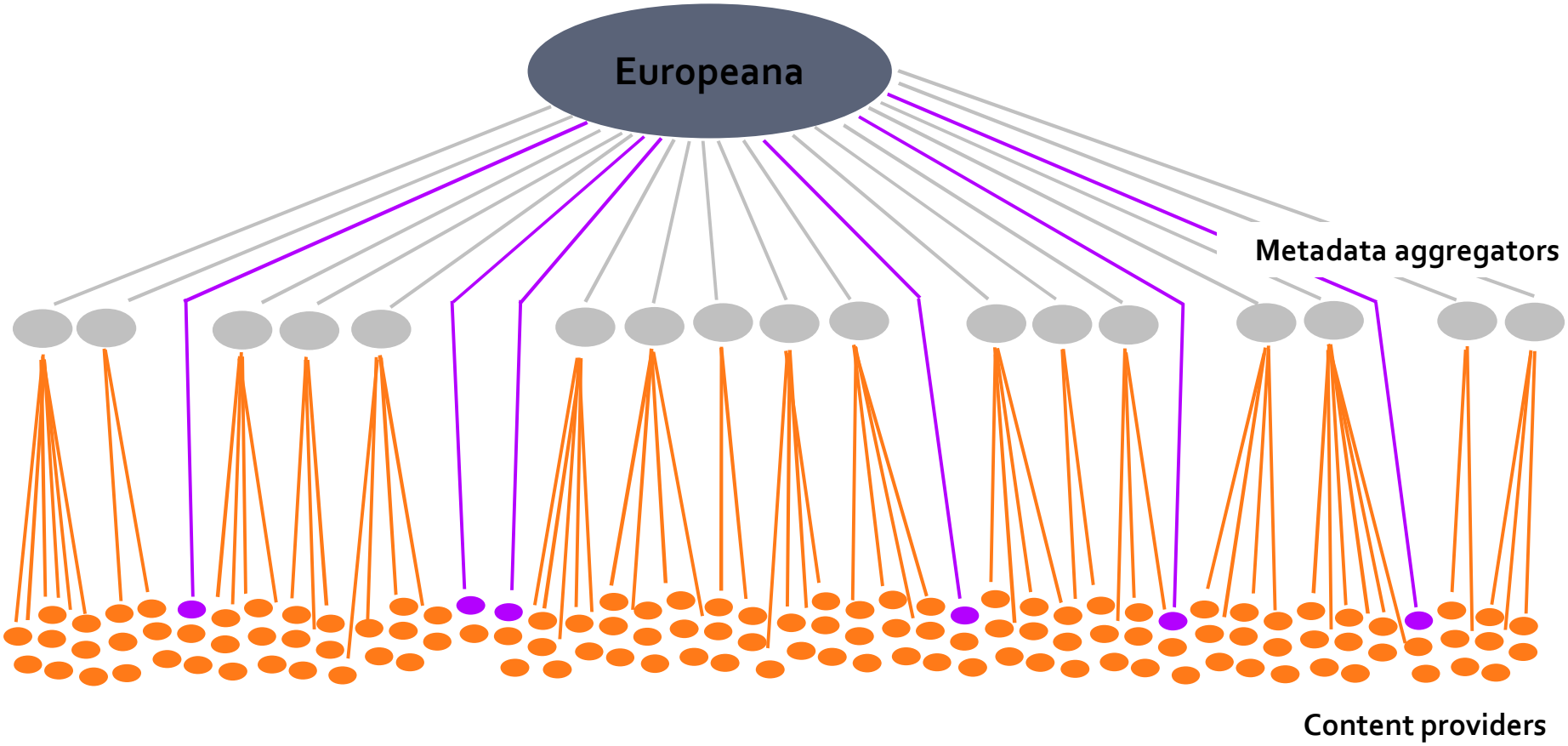
Schedule for the digital objects publication via Europeana

- EuropeanaLocal
  - May 2010 – 3 mlns
  - May 2011 – 10 mlns
- Europeana v1.0
  - July 2010 – „Rhine" – 10 mlns
  - April 2011 – „Danube"

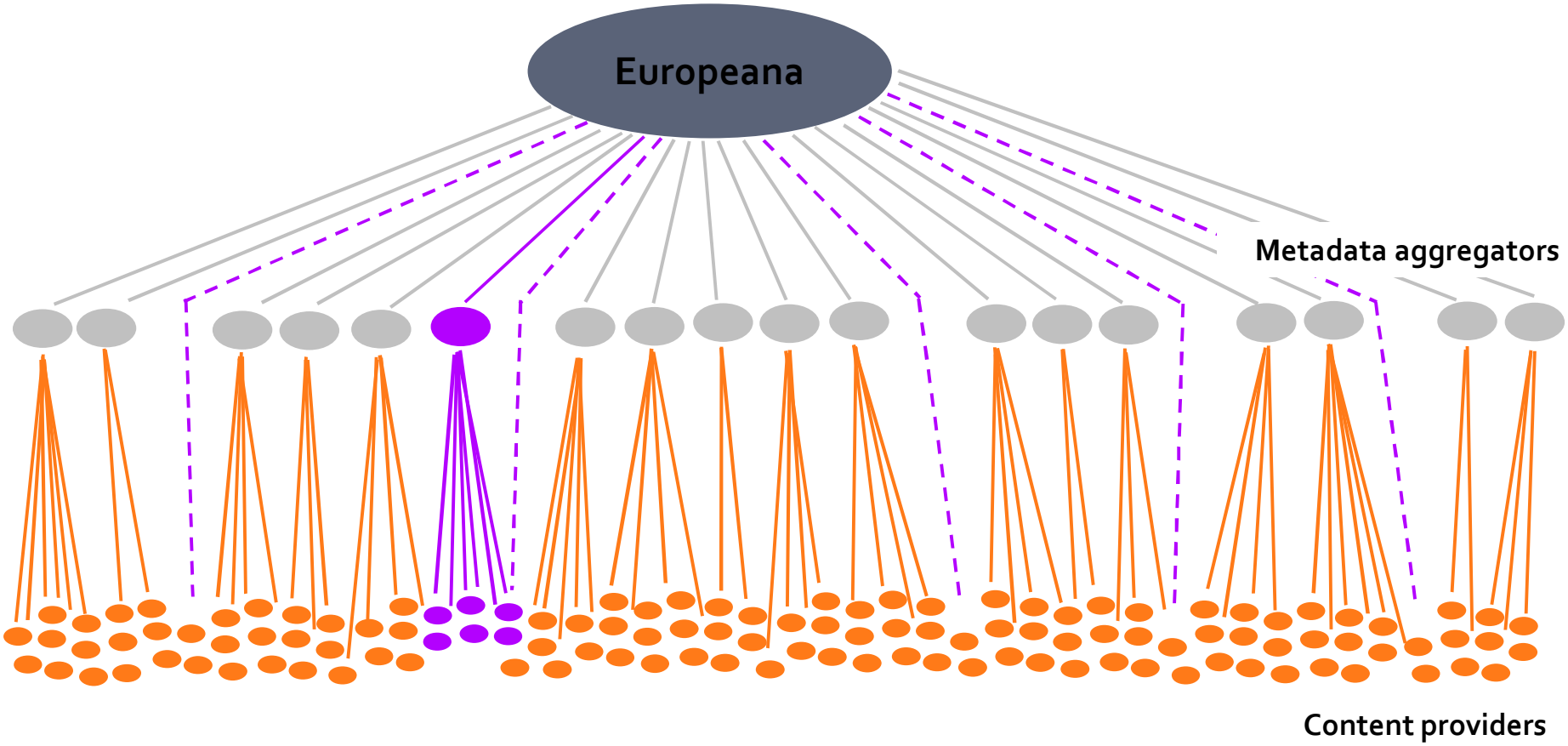


europeana
local

Present Europeana model for metadata aggregation



Target Europeana model for metadata aggregation



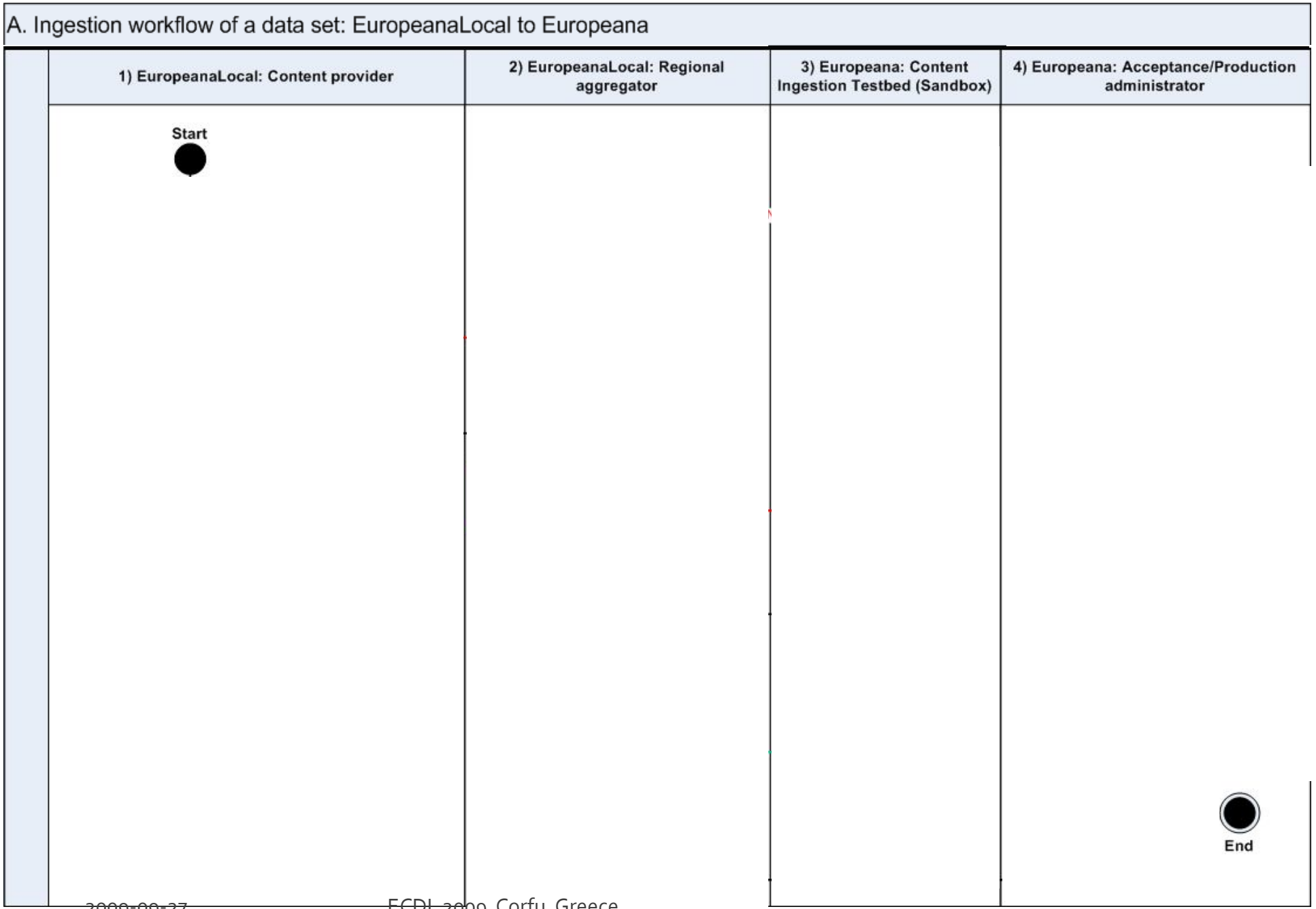
Metadata aggregators

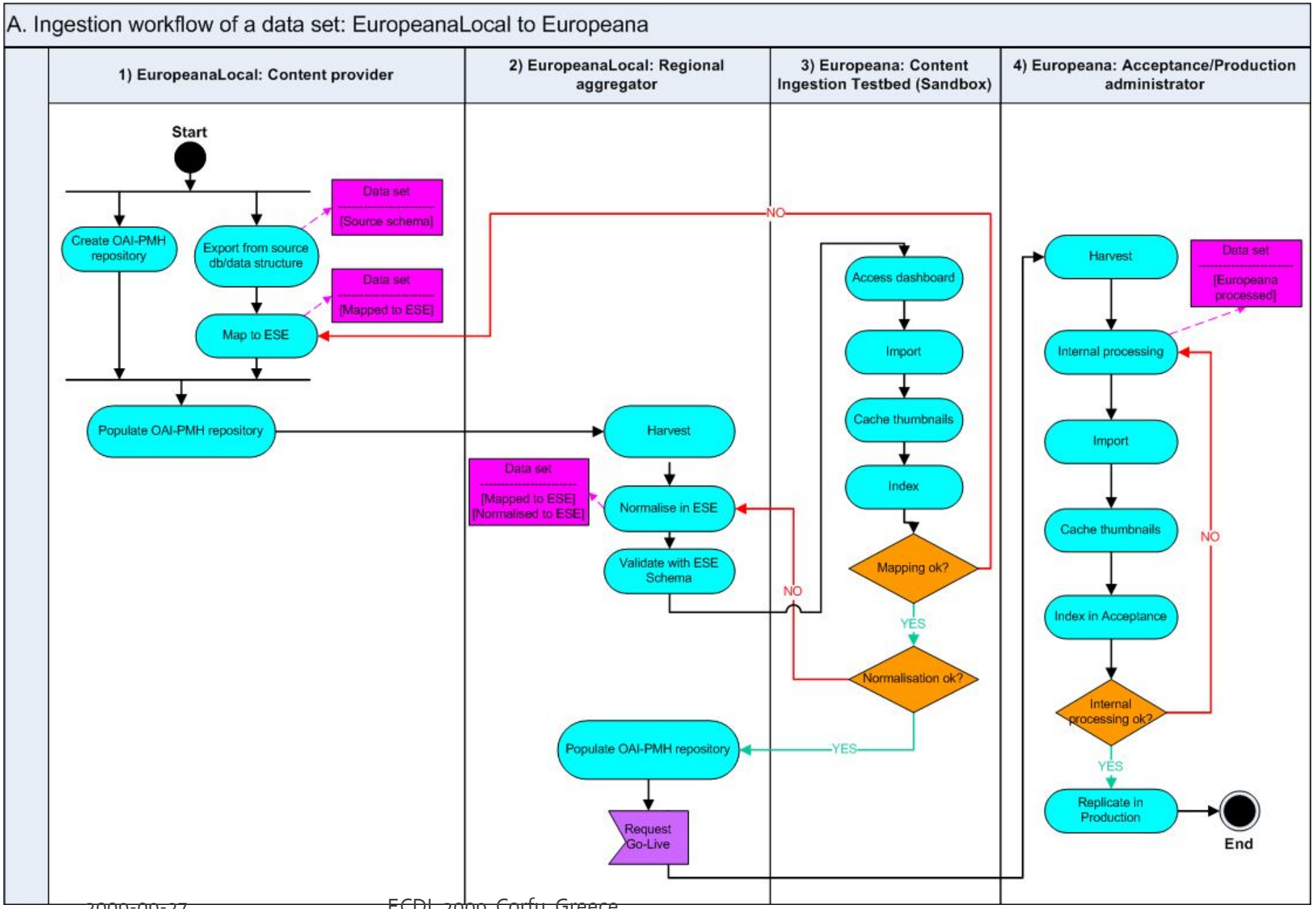
- According to the present version of Europeana Outline Functional Specification tasks for the aggregator are:
  1. To gather the information about content providers and their information systems
  2. To gather the metadata of objects that should be visible in Europeana
  3. To remove duplicates, clean-up the metadata, normalize it and enrich it
  4. To confirm the accessibility of digital objects
  5. To expose the aggregated metadata for Europeana via the OAI-PMH protocol

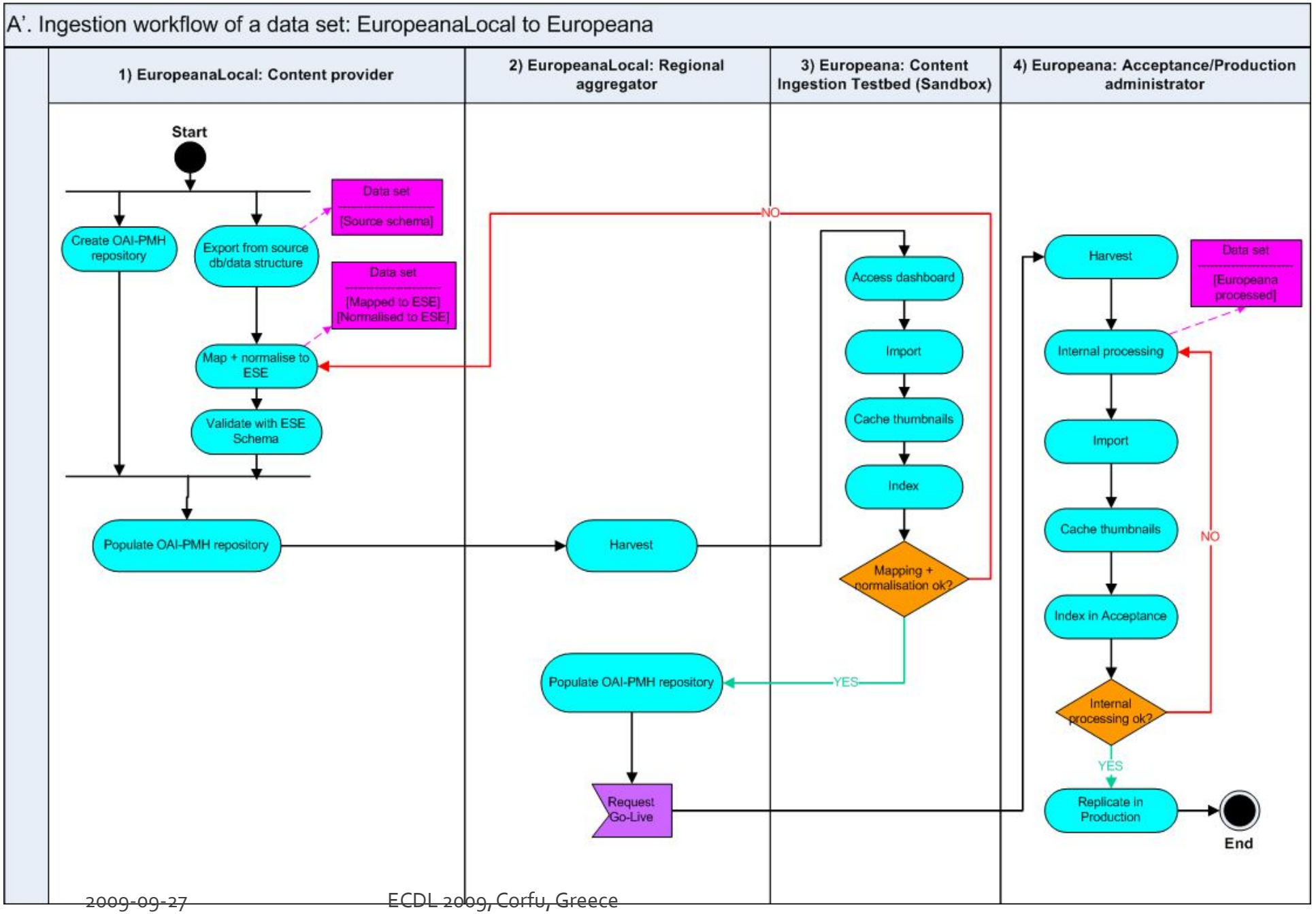
http://dev.europeana.eu/public_documents/EDLnet%20D2.5_Outline_Functional_Specifications20090301_version%201.7_consWithoutHistory_lossless.pdf

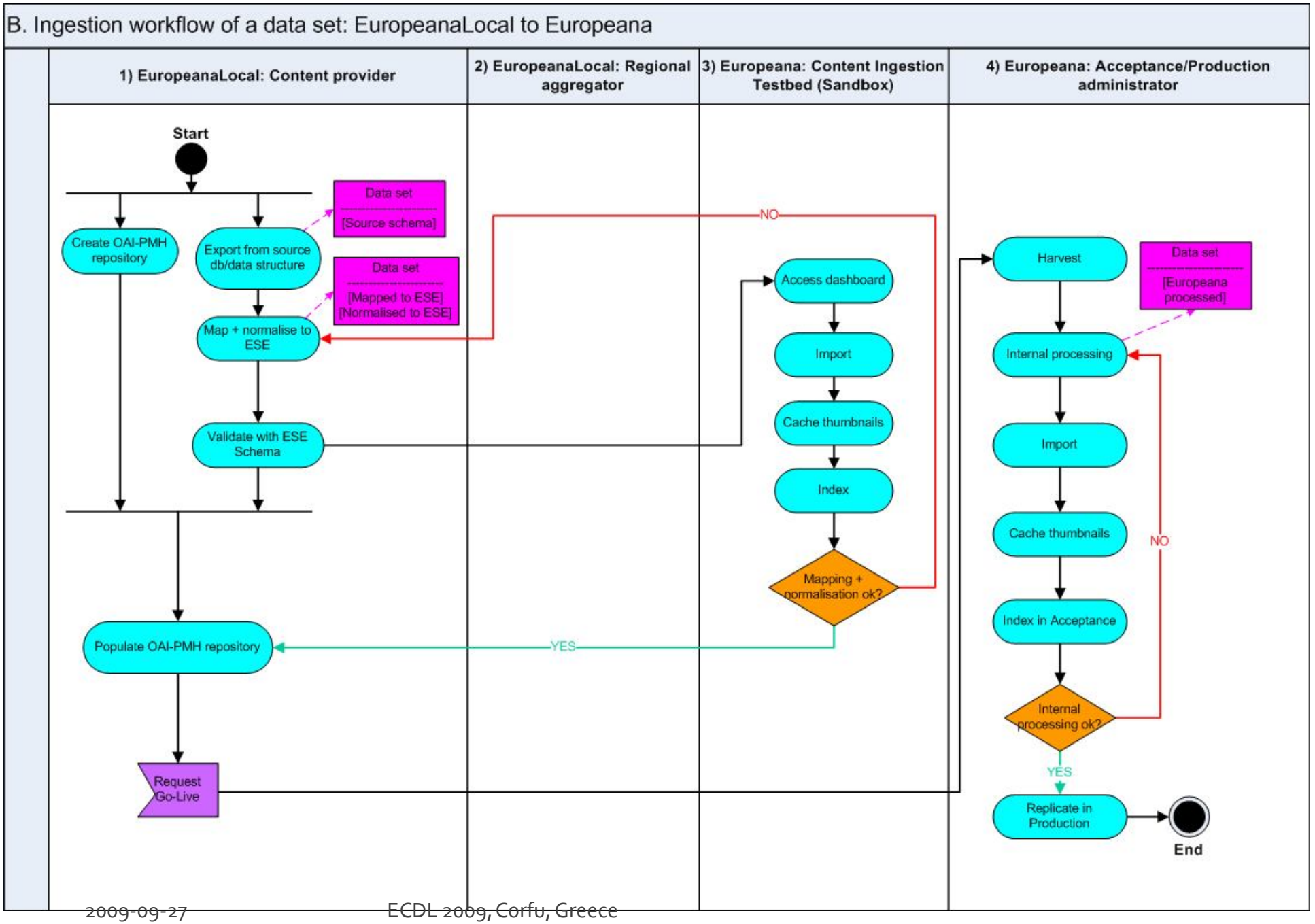
How to join Europeana?

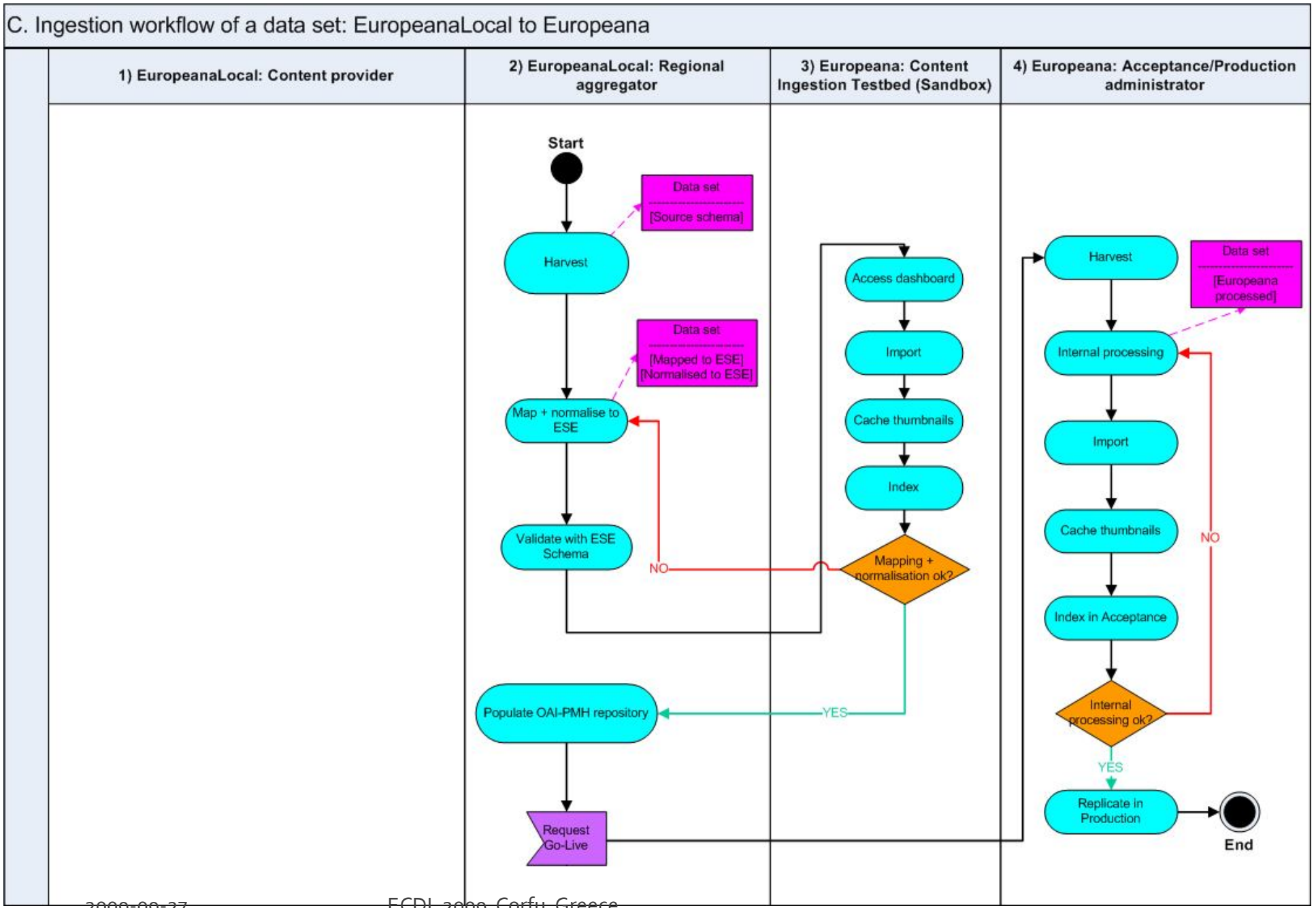
- Choose a metadata aggregator
- Map your metadata to Europeana Semantic Elements schema
- Normalize the metadata
- Test the metadata with Europeana
- Publish the metadata in the „production" version of Europeana


A. Ingestion workflow of a data set: EuropeanaLocal to Europeana
1) EuropeanaLocal: Content provider
2) EuropeanaLocal: Regional aggregator
3) Europeana: Content Ingestion Testbed (Sandbox)
4) Europeana: Acceptance/Production administrator
Start
End


A. Ingestion workflow of a data set: EuropeanaLocal to Europeana
1) EuropeanaLocal: Content provider
2) EuropeanaLocal: Regional aggregator
3) Europeana: Content Ingestion Testbed (Sandbox)
4) Europeana: Acceptance/Production administrator
Start
Create OAI-PMH repository
Export from source db/data structure
Data set
[Source schema]
Data set
[Mapped to ESE]
Map to ESE
Populate OAI-PMH repository
Harvest
Data set
[Mapped to ESE]
[Normalised to ESE]
Normalise in ESE
Validate with ESE Schema
Populate OAI-PMH repository
Request Go-Live
NO
Access dashboard
Import
Cache thumbnails
Index
Mapping ok?
NO
YES
Normalisation ok?
YES
Harvest
Data set
[Europeana processed]
Internal processing
Import
Cache thumbnails
Index in Acceptance
Internal processing ok?
NO
YES
Replicate in Production
End


A'. Ingestion workflow of a data set: EuropeanaLocal to Europeana
1) EuropeanaLocal: Content provider
2) EuropeanaLocal: Regional aggregator
3) Europeana: Content Ingestion Testbed (Sandbox)
4) Europeana: Acceptance/Production administrator
Start
Create OAI-PMH repository
Export from source db/data structure
Data set
[Source schema]
Data set
[Mapped to ESE]
[Normalised to ESE]
Map + normalise to ESE
Validate with ESE Schema
Populate OAI-PMH repository
Harvest
NO
Access dashboard
Import
Cache thumbnails
Index
Mapping + normalisation ok?
YES
Populate OAI-PMH repository
Request Go-Live
Harvest
Data set
[Europeana processed]
Internal processing
Import
Cache thumbnails
Index in Acceptance
Internal processing ok?
NO
YES
Replicate in Production
End


B. Ingestion workflow of a data set: EuropeanaLocal to Europeana
1) EuropeanaLocal: Content provider
2) EuropeanaLocal: Regional aggregator
3) Europeana: Content Ingestion Testbed (Sandbox)
4) Europeana: Acceptance/Production administrator
Start
Create OAI-PMH repository
Export from source db/data structure
Data set
[Source schema]
Data set
[Mapped to ESE]
[Normalised to ESE]
Map + normalise to ESE
Validate with ESE Schema
Populate OAI-PMH repository
Request Go-Live
NO
YES
Access dashboard
Import
Cache thumbnails
Index
Mapping + normalisation ok?
Harvest
Data set
[Europeana processed]
Internal processing
Import
Cache thumbnails
Index in Acceptance
Internal processing ok?
NO
YES
Replicate in Production
End


C. Ingestion workflow of a data set: EuropeanaLocal to Europeana
1) EuropeanaLocal: Content provider
2) EuropeanaLocal: Regional aggregator
3) Europeana: Content Ingestion Testbed (Sandbox)
4) Europeana: Acceptance/Production administrator
Start
Data set
[Source schema]
Harvest
Data set
[Mapped to ESE]
[Normalised to ESE]
Map + normalise to ESE
Validate with ESE Schema
NO
Access dashboard
Import
Cache thumbnails
Index
Mapping + normalisation ok?
Populate OAI-PMH repository
YES
Request Go-Live
Harvest
Data set
[Europeana processed]
Internal processing
Import
Cache thumbnails
Index in Acceptance
Internal processing ok?
NO
YES
Replicate in Production
End

Europeana Semantic Elements

- Metadata schema required by the Europeana
- Current version is 3.2, 07/08/2009
  - https://group.europeana.eu/c/document_library/get_file?uuid=c56f82a4-8191-42fa-9379-4d5ff8c4ff75&groupId=10602
- Metadata Mapping & Normalisation Guidelines for the Europeana Prototype
  - Version 1.2, 07/08/2009
  - https://group.europeana.eu/c/document_library/get_file?uuid=58e2b828-b5f3-4fe0-aa46-3dcbc0a2a1f0&groupId=10602

Europeana Semantic Elements

- ESE ver. 3.2 consists of:
  A. 15 Dublin Core elements
     + 22 Dublin Core qualifiers / terms
  B. 11 Europeana-specific elements
- Majority of elements from group A should be harvested from aggregated digital library
- Some of these elements may be extracted/mapped from other elements
  - It depends on the metadata standards used in particular digital library
- Majority (all?) of elements from group B may be extracted from A group elements or is obvious

Europeana Semantic Elements - Dublin Core

- Title
  - Alternative
- Creator
- Subject
- Description
  - Table of Contents
- Publisher
- Contributor
- Date
  - Created
  - Issued
- Type
- Format
  - Extent
  - Medium
- Identifier
- Source
- Language
- Relation
  - isVersionOf; hasVersion;
  - isReplacedBy; replaces;
  - isRequiredBy; requires;
  - isPartOf; hasPart;
  - isReferencedBy; references;
  - isFormatOf; hasFormat;
  - conformsTo
  - *isShownBy; isShownAt (Europeana)*
- Coverage
  - Spatial
  - Temporal
- Rights
- Provenance (DC Terms)

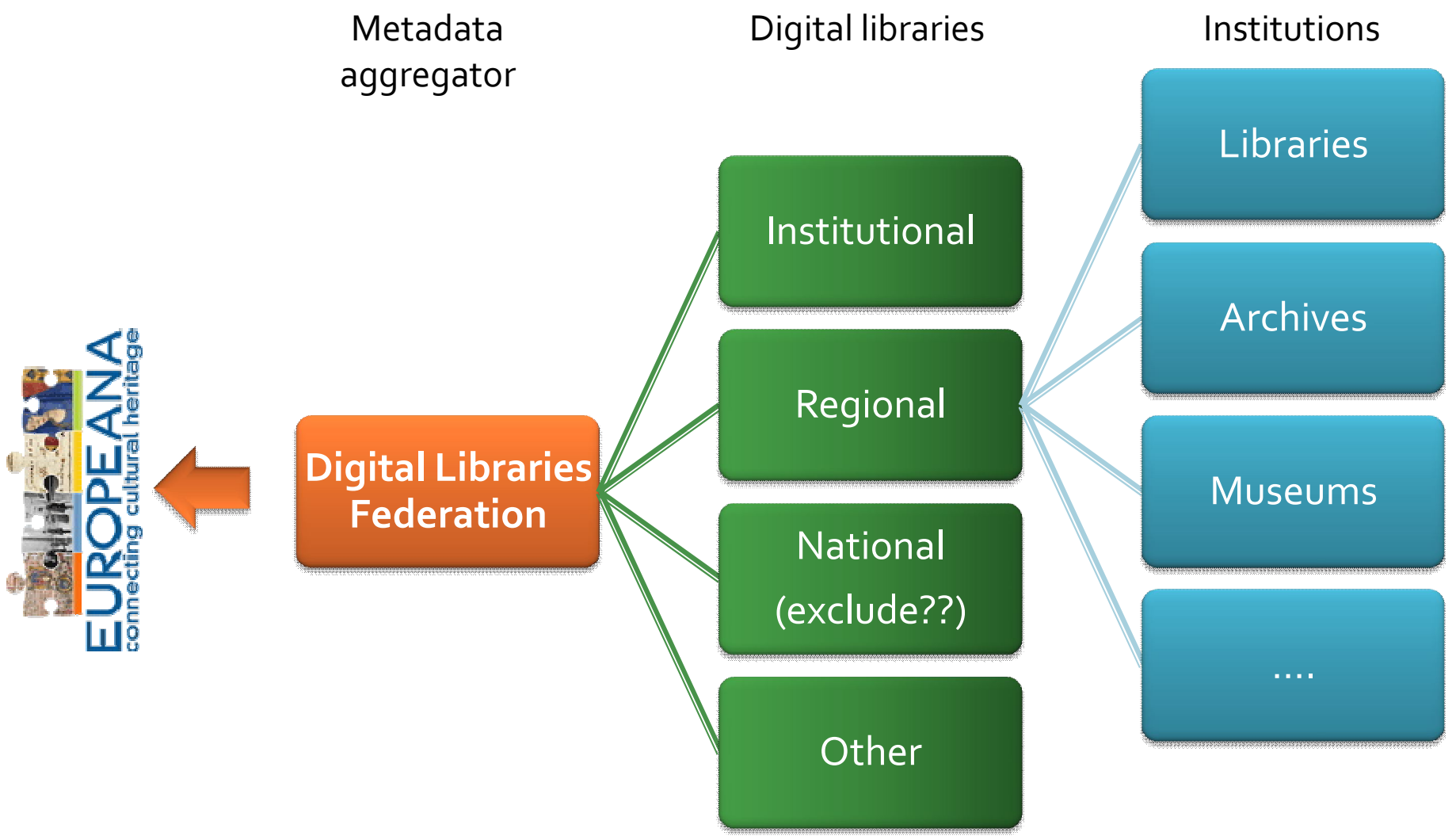
Europeana Semantic Elements - Europeana-specific elements

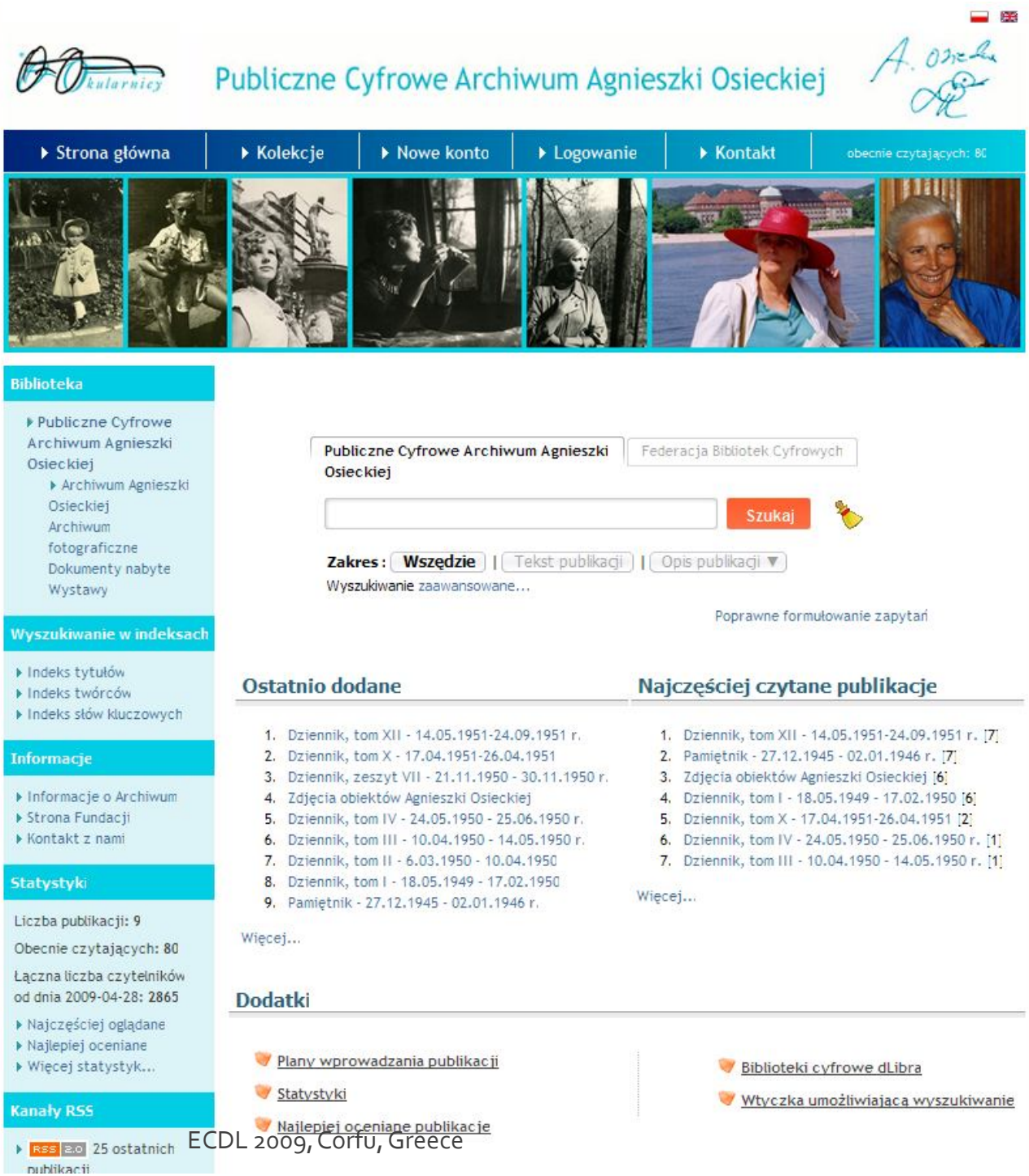
- **User tag** – user tags
- **Unstored** – everything that was not mapped to other fields
- **Object** – link to miniature/sample of an object
- **Language** – language of the country of the content provider
- **Provider** – provider of this object (aggregator)
- **Type** – object type (one of: Text, Image, Video, Sound)
- **URI** – unique identifier of the object
- **Year** – year related with the resource
- **Has Object** – is the field „Object" available
- **Country** – country of the content provider

Short summary about Europeana

- In the future Europeana has to be one of the main information points on European culture
  - Each European country should work on the highest possible representation in Europeana (currently ~50% objects comes from France)
- Because of the large scale of cooperation the basic organizational model will be based on aggregations
- Each content provider should decide which aggregator will provide its metadata to Europeana
  - Cooperation with several aggregators is also possible
- Metadata schema used by the Europeana is Europeana Semantic Elements
  - It is a Dublin Core qualified with 22 DC Terms and additionally 11 Europeana-specific elements
  - For the basic cooperation, the metadata in Dublin Core simple is enough but more metadata elements means better visibility in Europeana

Digital Libraries Federation as a metadata aggregator for Europeana




Publiczne Cyfrowe Archiwum Agnieszki Osieckiej
Strona główna
Kolekcje
Nowe konto
Logowanie
Kontakt
obecnie czytających: 80
Biblioteka
Publiczne Cyfrowe Archiwum Agnieszki Osieckiej
Archiwum Agnieszki Osieckiej
Archiwum fotograficzne
Dokumenty nabyte
Wystawy
Wyszukiwanie w indeksach
Indeks tytułów
Indeks twórców
Indeks słów kluczowych
Informacje
Informacje o Archiwum
Strona Fundacji
Kontakt z nami
Statystyki
Liczba publikacji: 9
Obecnie czytających: 80
Łączna liczba czytelników od dnia 2009-04-28: 2865
Najczęściej oglądane
Najlepiej oceniane
Więcej statystyk...
Kanały RSS
RSS 2.0 25 ostatnich publikacji
Publiczne Cyfrowe Archiwum Agnieszki Osieckiej
Federacja Bibliotek Cyfrowych
Szukaj
Zakres : Wszędzie | Tekst publikacji | Opis publikacji
Wyszukiwanie zaawansowane...
Poprawne formułowanie zapytań
Ostatnio dodane
1. Dziennik, tom XII - 14.05.1951-24.09.1951 r.
2. Dziennik, tom X - 17.04.1951-26.04.1951
3. Dziennik, zeszyt VII - 21.11.1950 - 30.11.1950 r.
4. Zdjęcia obiektów Agnieszki Osieckiej
5. Dziennik, tom IV - 24.05.1950 - 25.06.1950 r.
6. Dziennik, tom III - 10.04.1950 - 14.05.1950 r.
7. Dziennik, tom II - 6.03.1950 - 10.04.1950
8. Dziennik, tom I - 18.05.1949 - 17.02.1950
9. Pamiętnik - 27.12.1945 - 02.01.1946 r.
Więcej...
Najczęściej czytane publikacje
1. Dziennik, tom XII - 14.05.1951-24.09.1951 r. [7]
2. Pamiętnik - 27.12.1945 - 02.01.1946 r. [7]
3. Zdjęcia obiektów Agnieszki Osieckiej [6]
4. Dziennik, tom I - 18.05.1949 - 17.02.1950 [6]
5. Dziennik, tom X - 17.04.1951-26.04.1951 [2]
6. Dziennik, tom IV - 24.05.1950 - 25.06.1950 r. [1]
7. Dziennik, tom III - 10.04.1950 - 14.05.1950 r. [1]
Więcej...
Dodatki
Plany wprowadzania publikacji
Statystyki
Najlepiej oceniane publikacje
Biblioteki cyfrowe dLibra
Wtyczka umożliwiająca wyszukiwanie

Metadata aggregators

- According to the present version of Europeana Outline Functional Specification tasks for the aggregator are:
  1. To gather the information about content providers and their information systems
  2. To gather the metadata of objects that should be visible in Europeana
  3. To remove duplicates, clean-up the metadata, normalize it and enrich
  4. To confirm the accessibility of digital objects
  5. To expose the aggregated metadata for Europeana via the OAI-PMH protocol

http://dev.europeana.eu/public_documents/EDLnet%20D2.5_Outline_Functional_Specifications20090301_version%201.7_consWithoutHistory_lossless.pdf

Digital Libraries Federation as a metadata aggregator for Europeana

- To gather the information about content providers and their information systems
  - Database of Polish Digital Libraries in the DLF

Digital Libraries Federation as a metadata aggregator for Europeana

- To gather the metadata of objects that should be visible in Europeana
  - Done with the OAI-PMH
    - In most cases we require the OAI-PMH interface
    - In really special cases we can do it in different way (eg. Polish Internet Library)
  - Now we harvest only Dublin Core Simple
    - Works on new national metadata schema started in September 2009
    - Approximate time of development: 3 months
    - Approximate time of deployment: ???

Digital Libraries Federation as a metadata aggregator for Europeana

- **To remove duplicates**, clean-up the metadata, normalize it and enrich
  - Two types of duplication:
    - Duplicated metadata records describing the same digital object
    - Digital objects being a representation of the same physical object
      - Makes sense mostly in the context of libraries, where there may be several, practically identical editions of the same book
        - In museums and archives each object is unique
  - De-duplication in the DLF is based on the metadata comparison with some similarity threshold
    - Around 0.2% of aggregated objects makes the list of the „potential duplicates"
    - Similar mechanisms are used for the prevention of duplicated digitization

Digital Libraries Federation as a metadata aggregator for Europeana

- To remove duplicates, **clean-up the metadata, normalize it** and enrich
  - On the DLF level there are automatically built dictionaries on the basis of aggregated metadata
    - Separately for each metadata element
    - Separately for each metadata language
  - Differences between the metadata from various digital libraries have negative impact for the searching possibilities of the end-users
  - That is why the metadata normalization is so important
  - The basic analysis shows which elements are crucial and which should be easy to clean-up
    - The analysis was done in April 2009 on the metadata of 214 254 aggregated objects

Digital Libraries Federation as a metadata aggregator for Europeana

Element DC	Liczba unikalnych wartości	Liczba wystąpień tego elementu DC w opisach obiektów	Średnia liczba wyst. na poj. wartość
format	39	209 789	5 379,2
language	195	210 529	1 079,6
type	822	211 816	257,7
rights	1 192	246 093	206,5
coverage	66	2 390	36,2
publisher	18 002	310 764	17,3
contributor	12 979	83 464	6,4
subject	78 440	438 871	5,6
relation	9 292	48 319	5,2
date	47 581	209 589	4,4
identifier	6 426	27 666	4,3
description	43 657	180 391	4,1
source	16 996	52 506	3,1
creator	21 908	67 503	3,1
title	210 745	227 039	1,1

Digital Libraries Federation as a metadata aggregator for Europeana

- Format
  - In 99% of descriptions: MIME type(eg. text/html, image/x.djvu)
- Language
  - In most cases: ISO 639-2 (pol, ger, lat, fre etc.)
  - Sometimes one value „pol, ger" instead of „pol", „ger"
- Rights
  - Name of the institution which holds the original object
- Type
  - …

Digital Libraries Federation as a metadata aggregator for Europeana

Values for „Type” (top 20)	Number of objects with the value	% of aggregated objects	% of aggr. obj. (after clean-up)
czasopismo	44 709	20,9%	33,8%
gazeta	32 921	15,4%	31,3%
gazety	23 119	10,8%	
Czasopismo	20 965	9,8%	
książka	12 503	5,8%	
Gazeta	11 098	5,2%	
pocztówka	5 768	2,7%	
czasopisma	4 962	2,3%	
text	4 452	2,1%	
grafika	3 863	1,8%	
fotografia	3 596	1,7%	
artykuł z czasopisma	3 164	1,5%	2,6%
artykuł	2 455	1,1%	
Czasopisma	1 710	0,8%	
dzienniki urzędowe	1 516	0,7%	
stary druk	1 222	0,6%	1,1%
starodruk	1 221	0,6%	
rysunek	1 094	0,5%	
rękopis	1 062	0,5%	
mapa	1 028	0,5%	
Sum		**85,1%**	**68,9%**

Digital Libraries Federation as a metadata aggregator for Europeana

- To remove duplicates, clean-up the metadata, normalize it and **enrich**
  - Basic enrichment can be the creation of the Europeana specific metadata elements from other Dublin Core fields

Digital Libraries Federation as a metadata aggregator for Europeana

- Europeana specific elements
  - isShownBy, isShownAt
    - Links to objects used in Europeana interface
  - unstored
    - Place for everything that cannot be mapped to ESE
  - object – Link to the miniature/sample of the digital objects
    - Creation of such link can be sometimes automated
      - http://www.wbc.poznan.pl/dlibra/docmetadata?id=2752
      - http://www.wbc.poznan.pl/Content/2752
      - http://www.wbc.poznan.pl/image/edition/2752
  - hasObject
    - true or false – shows if the „object" field is present

Digital Libraries Federation as a metadata aggregator for Europeana

- Europeana specific elements
  - provider
    - Name of the content provider (aggregator)
  - country
    - Country of the content provider (ISO 3166)
  - language
    - Official language in the country of the content provider (ISO 639-1)
  - uri
    - Unique resource identifier

Digital Libraries Federation as a metadata aggregator for Europeana

- Europeana specific elements
  - type
    - One of following values TEXT, IMAGE, SOUND, VIDEO
    - Can be in most cases chosen automatically
      - On the basis of dc:type i dc:format
  - userTag
    - Tags created by users (of Europeana??)
  - year
    - 4 digit number (???) in the Gregorian calendar used for time navigation
    - In many cases can be automatically extracted from the dc:date

What about vertical services?

- Europeana wants to aggregate all publicly available digital content relevant to the term „European cultural and scientific heritage"
- What about vertical services based on a large scale aggregation?
  - There is a need to enable precise selective harvesting of aggregated metadata

Example scenario: Thematic portal built on top of distributed OAI-PMH repositories

- How to obtain the metadata?
  - Solution 1: Harvest all records from repositories, decide what records are useful
    - A lot of useless data is harvested and processed
  - Solution 2: Harvest only specific sets of items matching the theme of the portal
    - Each harvested repository must define a set / sets matching the theme of the portal – practically impossible
  - Solution 3: DIY variant of scenario 2 – define a set containing items matching the theme of the portal and harvest it
    - Not supported in the OAI-PMH protocol

Proposed OAI-PMH extension: dynamic sets

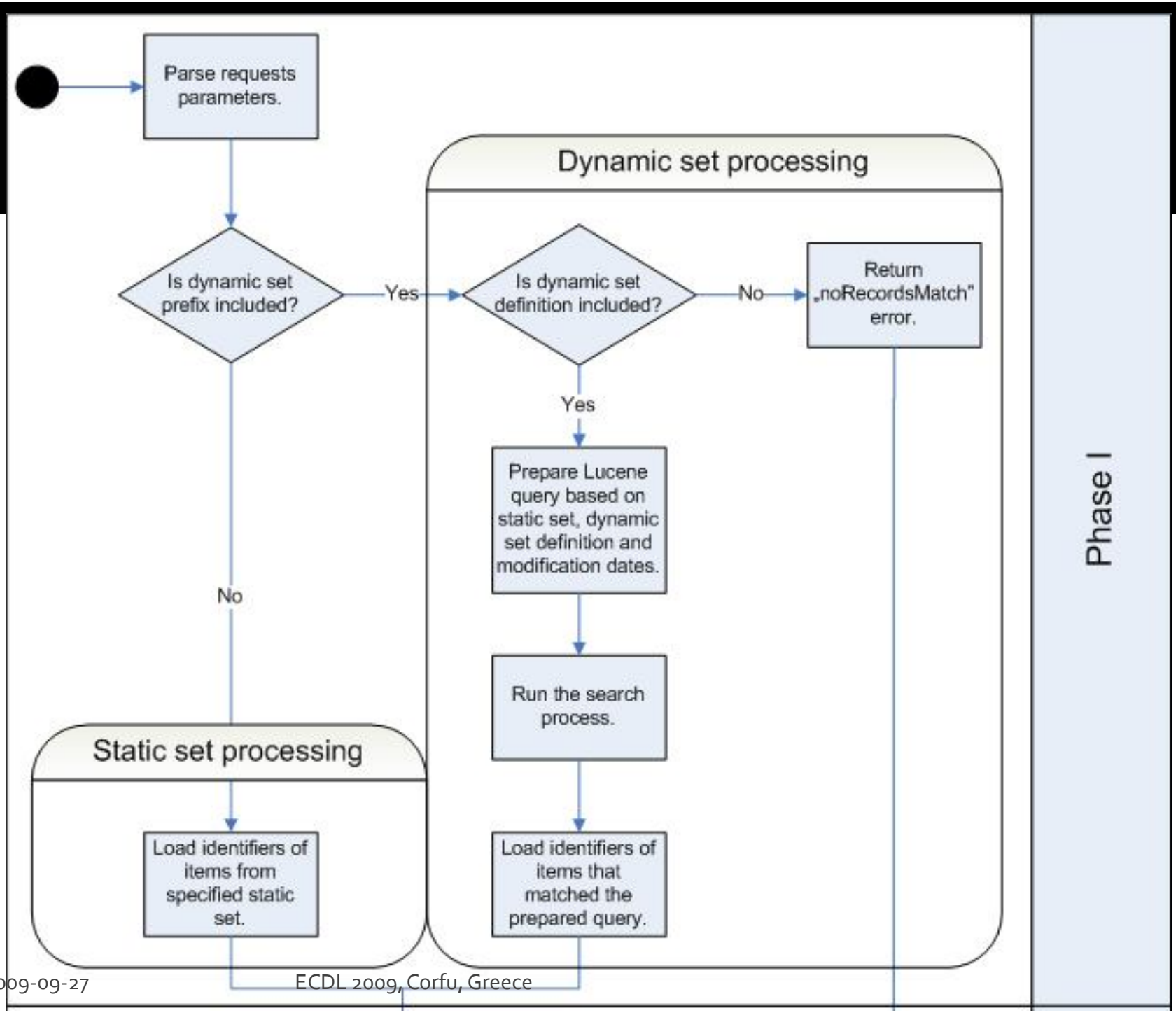
- Dynamic sets – specification
  - Sets defined by repository users
  - Contain items that matched dynamic set definition sent by the user
  - The definition is in fact a CQL query encoded into a set name
    - CQL – Contextual Query Language
      - Part of SRU protocol specification – used in integrated library systems as a replacement for the z39.50 protocol to obtain bibliographic descriptions
      - Allows to define simple and complex queries
      - Compatible with any metadata schema
      - Example: dc.creator = "Albert Einstein"

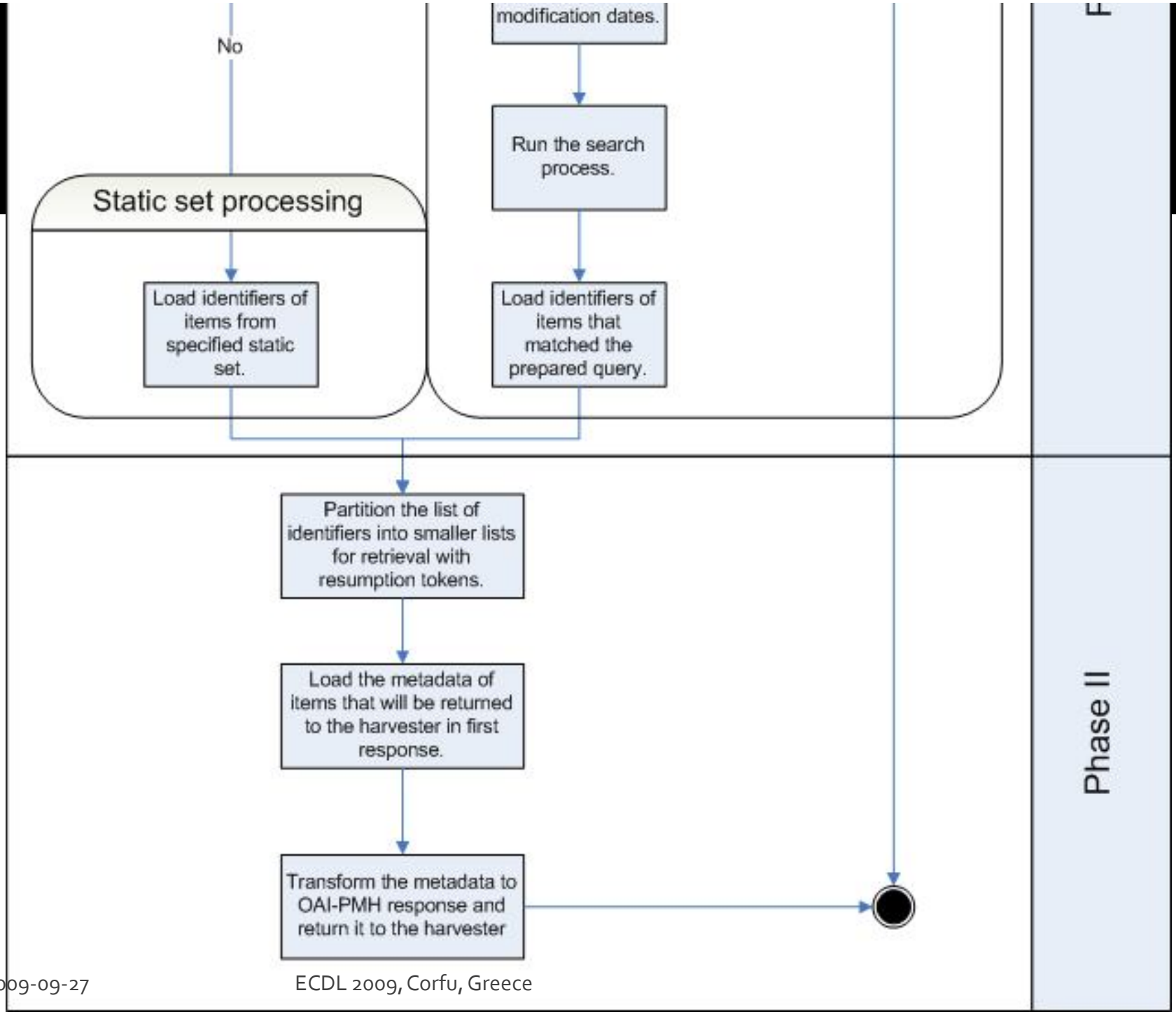
Proposed OAI-PMH extension: dynamic sets

- Dynamic sets – OAI-PMH protocol compatibility
  - Harvester side
    - If a harvester does not supports dynamic sets, it will be still able to harvest the repository supporting such sets
  - Repository side
    - If a repository does not supports dynamic sets, it still may be harvested by a harvester supporting such sets
    - The repository extended with dynamic sets should be compatible with OAI-PMH validators

Proposed OAI-PMH extension: dynamic sets

- Dynamic sets – implementation
  - Harvester side
    - Prepare the support for OAI-PMH set harvesting
    - Analyze the nature of metadata in particular repository and prepare proper dynamic set definition to use during harvesting
  - Repository side
    - Modify the harvesting requests processing to support the definition of dynamic sets
    - This may be based on the search mechanism already implemented in the majority of repositories – in such case the support for CQL queries must be assured


Parse requests parameters.
Dynamic set processing
Is dynamic set prefix included?
Yes
Is dynamic set definition included?
No
Return „noRecordsMatch" error.
Yes
Prepare Lucene query based on static set, dynamic set definition and modification dates.
Run the search process.
Load identifiers of items that matched the prepared query.
No
Static set processing
Load identifiers of items from specified static set.
Phase I


No
modification dates.
Run the search process.
Static set processing
Load identifiers of items from specified static set.
Load identifiers of items that matched the prepared query.
Partition the list of identifiers into smaller lists for retrieval with resumption tokens.
Load the metadata of items that will be returned to the harvester in first response.
Transform the metadata to OAI-PMH response and return it to the harvester
Phase II

Prototype implementation and tests

- Prototype implementation of the OAI-PMH extension in the Digital Libraries Federation software
- Test harvests
  - dc.language eng – publications written in English
  - dc.language ger – publications written in German
  - dc.type podręcznik (handbook) – publications of type handbook
  - dc.type rozprawa (thesis) – publications of type thesis
  - dc.type czasopismo (magazine) – publications of type magazine
  - dc.type gazeta (newspaper) – publications of type newspaper
  - dc.subject pedagogika (pedagogy) – publications about pedagogy
  - dc.subject chemia (chemistry) – publications about chemistry

Tests results

Query	Harvested number of		Harvested % of overall number of	
	repositories	records	repositories	records
none (all records)	*16*	*93681*	*100,00%*	*100,00%*
dc.language *eng*	13	626	81,25%	0,67%
dc.language *ger*	12	10357	75,00%	11,06%
dc.type *podręcznik* (handbook)	4	104	25,00%	0,11%
dc.type *rozprawa* (thesis)	5	199	31,25%	0,21%
dc.type *czasopismo* (magazine)	16	28163	100,00%	30,06%
dc.type *gazeta* (newspaper)	4	33793	25,00%	36,07%
dc.subject *pedagogika* (pedagogy)	8	130	50,00%	0,14%
dc.subject *chemia* (chemistry)	8	715	50,00%	0,76%
dc.subject	8	2759	50,00%	2,95%

Current usage

- eContentPlus ENRICH Project (PSNC is a participant)
  - Started in December 2007
  - The aim is to built a virtual European repository of manuscripts
  - The metadata about the manuscripts is harvested from multiple European repositories
  - Harvests metadata of manuscripts from several Polish digital libraries

Summary

- Present directions of the development of European data infrastructure are extensively using large scale metadata aggregation
- Semantic interoperability and selective harvesting are one of the crucial issues in this approach
- What we have presented today are the experiences from the development of the Polish digital libraries infrastructure
- We hope that you will find it useful when facing the same task in your country, region or domain