Inclusion of a new country in ARROW: what does it mean in practice?
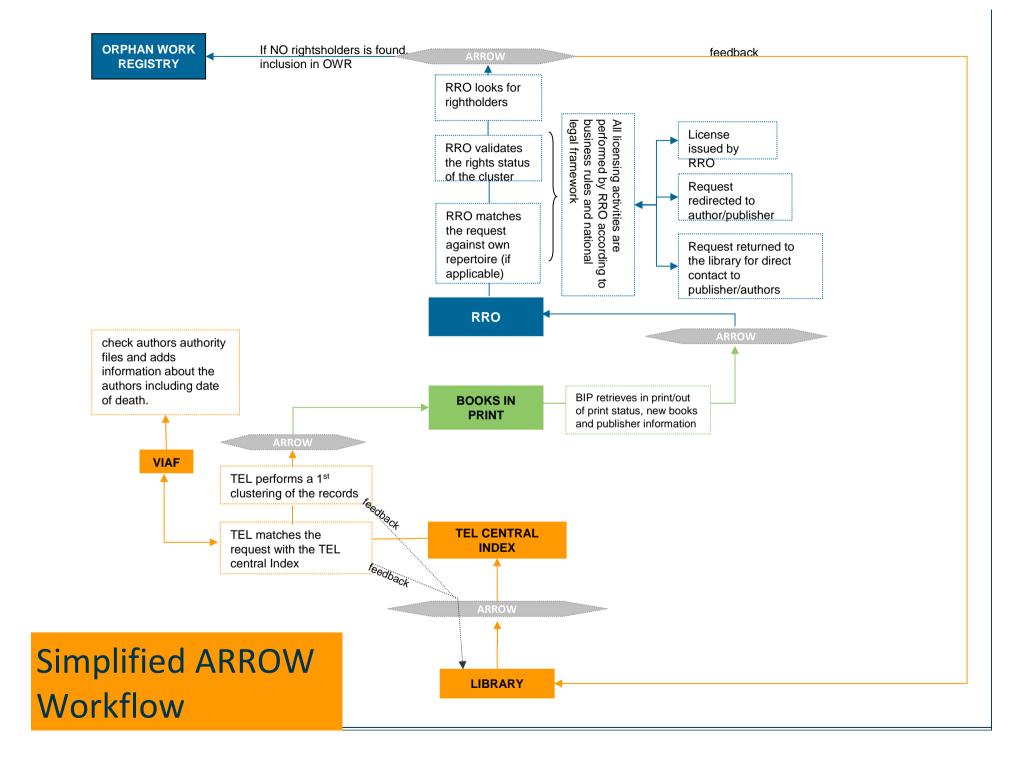
ARROW Plus kick off meeting - Bologna, 29th June 2011

www.arrow-net.eu

ORPHAN WORK REGISTRY

If NO rightsholders is found, inclusion in OWR

ARROW

feedback

RRO looks for rightholders

RRO validates the rights status of the cluster

RRO matches the request against own repertoire (if applicable)

All licensing activities are performed by RRO according to business rules and national legal framework

License issued by RRO

Request redirected to author/publisher

Request returned to the library for direct contact to publisher/authors

RRO

ARROW

check authors authority files and adds information about the authors including date of death.

BOOKS IN PRINT

BIP retrieves in print/out of print status, new books and publisher information

VIAF

ARROW

TEL performs a 1st clustering of the records

feedback

TEL matches the request with the TEL central Index

TEL CENTRAL INDEX

feedback

ARROW

LIBRARY

Simplified ARROW Workflow

Including new countries

- In each country a minimum set of requirements needs to be met in the book data infrastructure for ARROW inclusion:
  - The national library catalogue should be in TEL
  - Books in print exists
  - Reliable RRO database exists
- How to move towards countries where one or more of these elements lack?
  - We need to create the data infrastructure where it does not exist
  - Or improve the existing data infrastructure where it is not able to fit the requirements

Let's see how we need to work in both cases

Assess each country and gather requirements/1

WP3

- What needs to be done:
  - For each country, understand what is the situation in the Library domain, the BIP domain, the RRO domain
    - Identification and selection of relevant stakeholders to be involved
    - Assessment of the existing book data infrastructure
    - Assessment of the need to improve existing infrastructure or create new infrastructure from scratch
    - Specification of technical requirements for integration of infrastructure in the ARROW system

  - Countries priority list for inclusion
    - According to previously identified characteristics

Assess each country and gather requirements/2

- What needs to be done:
  - Methodology:
    - Fact finding: questionnaires and ad hoc colloquia on BIP and RRO registries, when needed
    - Identification and activation of national reference partners and of national groups of relevant stakeholders
    - Organisation of national stakeholder meetings (general and individual)
    - Preparation and refinement of requirements table; filling of requirements table with information from stakeholders to support WP5 and WP4

Assess each country and gather requirements/3

- What do the National contact points and National groups need to do?
    - provide the contact details of all the relevant stakeholders
    - assist the WP leader in contacting the relevant stakeholders (NCP)
    - provide information about the existing sources of bibliographic data (national bibliography, BIP, RRO repertoires and alternatives), their characteristics, the requirements for those to be included in the Arrow workflow
    - where such sources are lacking, illustrate any potential plans to create them and the requirements for this to be accomplished
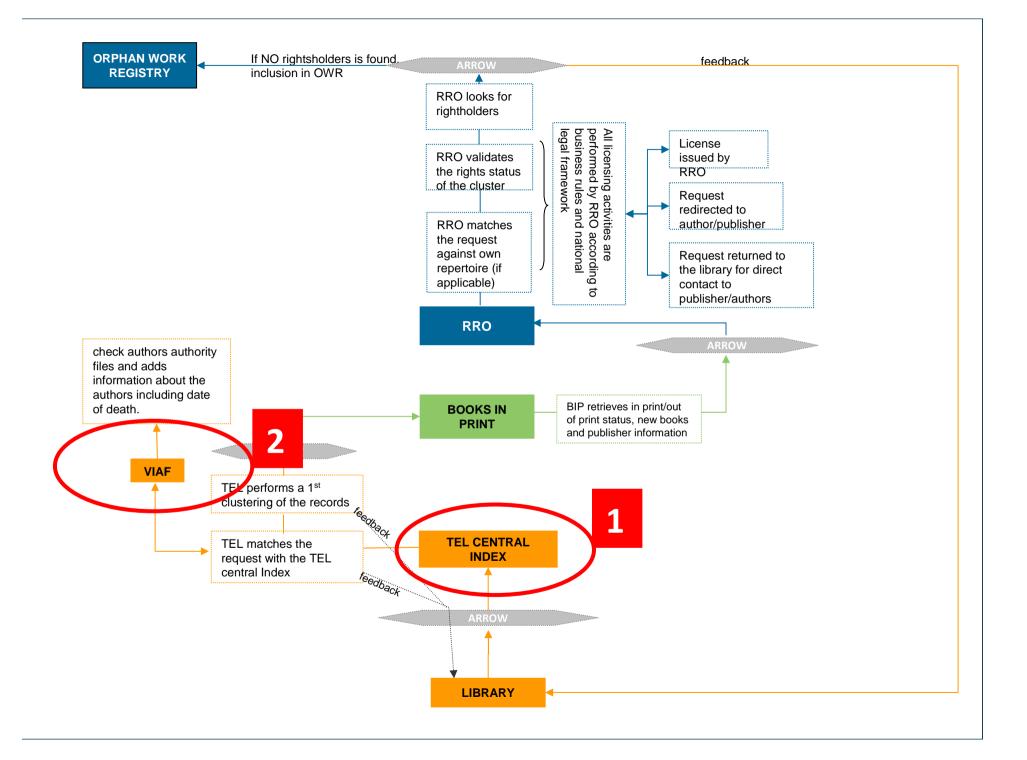   - take part to national meetings and ad hoc colloquia when requested

The importance of National groupings/1

WP3

- Because in each country the project needs at least 3 types of participants: national library, BIP and RRO, a leader must be identified at the national level to coordinate the work

  - to organize regular meetings between national participants (one per month in France), and prepare the minutes;
  - to circulate and explain information coming from the project to the national participants
  - to report to FEP, AIE and to the other external partners on the work done at the national level or on the national expectations
  - to coordinate the work between the national participants in respect of the schedule : to be sure that the 3 types of participants are ready for the D Day!
  - to circulate and explain information coming from external existing networks (TEL or CENL, FEP, BIPs network, RROs network, etc.)
  - to involve all the team in the preparation of the international meetings, specially the reviews by the European Commission

"BnF on the ARROW French National Group"

The importance of National groupings/2

- Involve in the national meetings people who represent the national administrative supervision: Ministry of Culture and Communication was represented in France, together with the National publisher association.

- These regular national meetings will create sound relations between the national participants, especially if there is no experience of working together. This will appear very valuable especially for the test period when the 3 main participants have to define common use cases

"BnF on the ARROW French National Group"

ORPHAN WORK REGISTRY

If NO rightsholders is found, inclusion in OWR

ARROW

feedback

RRO looks for rightholders

RRO validates the rights status of the cluster

RRO matches the request against own repertoire (if applicable)

All licensing activities are performed by RRO according to business rules and national legal framework

License issued by RRO

Request redirected to author/publisher

Request returned to the library for direct contact to publisher/authors

RRO

ARROW

check authors authority files and adds information about the authors including date of death.

BOOKS IN PRINT

BIP retrieves in print/out of print status, new books and publisher information

2

VIAF

TEL performs a 1st clustering of the records

feedback

1

TEL matches the request with the TEL central Index

TEL CENTRAL INDEX

feedback

ARROW

LIBRARY

Library Domain requirements

☐ **That National Bibliography Records are or can be made available in TEL**

  ◻ **STATUS:** (almost) essential

  ◻ **REASON:** to allow TEL Processing: matching, work extraction, clustering

☐ **That the National Name Authority File has been merged with VIAF**

  ◻ **STATUS:** nice to have

  ◻ **REASON**: to add authoritative data about authors and contributors, forms of the names, dates of birth and death, nationality.

How catalogues are made available for ARROW in The European Library

WP4

- Four general steps
  - Ingesting the catalogue
  - Work extraction from MARC records
  - Indexing for searching, matching and clustering
  - Testing

Ingesting the catalogue

WP4

- Supported metadata formats (for ingestion):
  - UNIMARC or MARC21 in either ISO2709, MARCXML or MarcXchange
  - Support for other MARC formats is expected to be necessary in ARROW Plus
    - For example Austria and Hungary

- Transfer mechanisms
  - FTP server (Libraries' own, or TEL's)
  - OAI-PMH
  - DVDs, flashdisks, etc…
  - An automated method like FTP or OAI-PMH is recommended

Recommendations for the national libraries

WP4

- Include the identifier in the library's authority file for authors and contributors

  - Subfield $0 of fields 100, 110, 700, 710, …

  - It will allow linking to VIAF (if library participates in VIAF)

  - It will allow the matching and clustering to know if two "John Smiths" are the same person or not

- Automatic of semi-automatic transfer mechanisms

  - Libraries may choose the most practical transfer method at this time

  - …but for an operational service, a method like FTP or OAI-PMH would be recommended

Work extraction from MARC records

WP4

- Most components of the TEL system are based on data structures for representing Works instead of MARC
  - This minimizes dependencies of the metadata formats of the catalogues
- After ingest, all data representing the Work is extracted from the MARC record
  - The data structure for Works is inspired on ISTC metadata
- Work extraction for MARC21 is documented in D6.1 of the ARROW project

Indexing for searching, matching and clustering

WP4

- The European Library Metadata Repository
  - Stores the extracted work metadata
  - Stores the MARC record
    - Exchange of records to ARROW is always in MARC21

- The Similarity Search Engine
  - Allows similarity queries on the work metadata
  - Supports the matching and clustering of manifestations and works
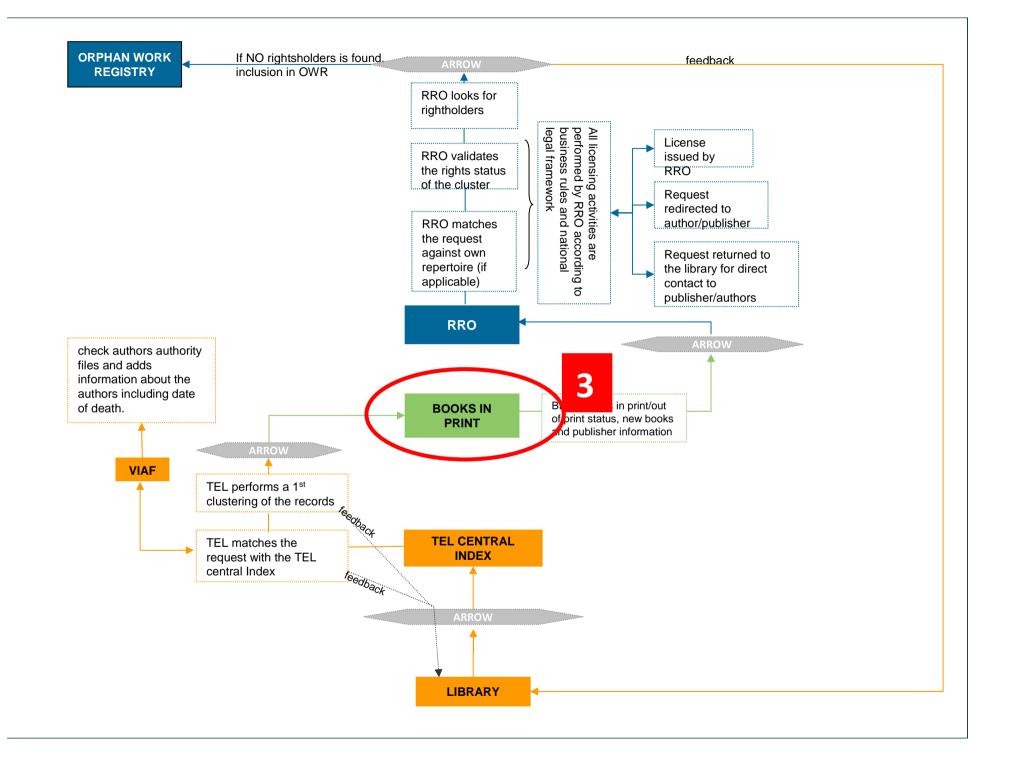
Testing the catalogue

- Testing is based on a ~20 record sample chosen by the national library

  - It should be the same sample used to test the ARROW workflow

- An ideal sample is heterogeneous. It should have:

  - Coverage of different time periods

  - Mainly books published in the country

  - Two books published in other countries already in ARROW

  - Books with ISBN and books without ISBN

  - Both books by a single author and books by multiple authors

  - Books that are translations

  - Some books with many editions

- Work extraction, MARC21 conversion, Clustering and Matching results are checked with the national library

Country adjustments

- The TEL system may work differently for each country

- Based on the rights clearance requirements of each country

- Characteristics of the data in the catalogues may also require adaptations

  - Different cataloguing practices
  - Different data characteristics
  - Different data quality issues

ORPHAN WORK REGISTRY

If NO rightsholders is found, inclusion in OWR

ARROW

feedback

RRO looks for rightholders

RRO validates the rights status of the cluster

RRO matches the request against own repertoire (if applicable)

All licensing activities are performed by RRO according to business rules and national legal framework

License issued by RRO

Request redirected to author/publisher

Request returned to the library for direct contact to publisher/authors

RRO

ARROW

3

check authors authority files and adds information about the authors including date of death.

BOOKS IN PRINT

Books in print/out of print status, new books and publisher information

ARROW

VIAF

TEL performs a 1st clustering of the records

feedback

TEL matches the request with the TEL central Index

TEL CENTRAL INDEX

feedback

ARROW

LIBRARY

BIP Domain requirements

☐ **That a comprehensive and accurate Books in Print database exists for the territory and can be queried as part of the ARROW workflow**

　　◻ **STATUS:** (almost) essential

　　◻ **REASON:** to retrieve information of commercial availability of the work and all its manifestations

☐ **That there is a web-services query interface to Books in Print data that is compliant with ARROW system standards**

　　◻ **STATUS:** highly desirable

　　◻ **REASON:** to allow machine to machine interaction and automatic workflow

Integrate an existing BIP/1

WP4

- Two general approaches:
  - BIP has already in place its own synchronous B2B query interface
  - BIP decides to implement a B2B query interface compliant with ARROW system standard

- Other scenarios may require an ad hoc approach

Integrate an existing BIP/2 Approach 1

WP4

- □ CINECA has to:

    - ◘ Set up B2B client (Connector)
        - ■ Study service query interface
        - ■ Implement/Test the B2B client (Semantic tests may be necessary)

    - ◘ Integrate Connector in ARROW workflow (WF)
        - ■ Implement test mapping of BIP query output metadata format (if not ONIX) in ARROW ShortDescription

    - ◘ Extend/Customize and Test business logic (WF)
        - ■ Build necessary data structure to query BIP
        - ■ Build M6R from gathered results (applying filtering, duplicate check etc)

    - ◘ Test whole WF. Semantic test must be performed on the gathered results (not under CINECA responsibility).

- □ BIP has to:

    - ◘ Provide support during the development/test phase
    - ◘ Customize/refine query parameters if necessary

Integrate an existing BIP/3 Approach 2

WP4

- BIP has to:
  - Provide BIP Query Web Service (M6Q)
    - Set up the B2B interface in order to receive M6Q
    - Test B2B interface
  - Provide BIP Response ARROW Web Service Client
    - Implement/Test Web Service client in order to send to ARROW system the query results (M6R)
  - Design/Implement business logic
    - Elaborate ARROW M6Q
    - Send results to ARROW WS
    - Retry mechanisms
- CINECA has to:
  - Test BIP Query Web Service
  - Implement/Test mapping of BPP query output metadata format in ARROW ShortDescription
  - Integrate new Connector in ARROW Workflow
  - Test whole workflow
  - Provide support to BIP during development/test phases

Set up a new BIP/1

- What needs to be done
  - Gather and Analyse BIP system requirements (CINECA/MVB)
  - Create a data-model that fits local and international needs / that ensures interoperability, e.g. based on ONIX
  - Establish Services for publishers
    - Web interface for adding new titles and editing existing title
    - Possibility to upload Multi-Media-Objects for every title (MM)
    - Provide Quality Assurance Tools (QA)
    - Title List Import (ONIX, XLS, CSV)
    - Title List Export (ONIX, XLS, CSV)
  - Establish Services for booksellers
    - Web interface for searching titles
    - Meta Data SOAP Webservice
    - Title List Export (ONIX, XLS, CSV)
    - Ordering Feature (Optional)

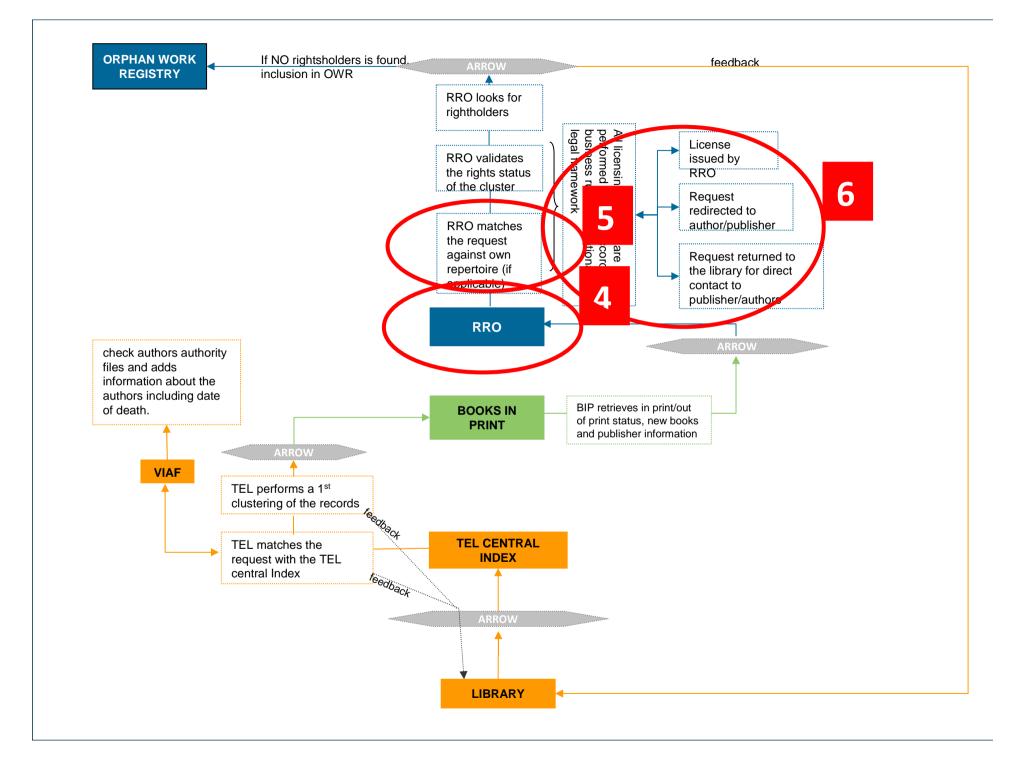# Set up a new BIP/2

- What needs to be done
    - Establish Backend systems
        - User Management
        - Monitor Data Processes
        - Access to BIP catalogue for editorial department
    - Set up of BIP system
        - Design BIP system architecture (CINECA/MVB)
        - Implementation/Test of BIP System
    - BIP feeding
        - Analysis of existing resources
        - Analysis and Realization of an import tool from existing source
        - Data mapping
    - Analysis of BIP integration in ARROW Workflow

Set up a new BIP/3

WP5

- What do partners in the country need to do
  - Intensive participation in the requirements-gathering process
  - Close collaboration with the WP5 leaders during testing-process the new BIP registry
  - Giving the WP5 leaders consolidated feedback about how the implementation of the new BIP registry will be accepted by the users in the local market (publishers, booksellers)

ORPHAN WORK REGISTRY

If NO rightsholders is found, inclusion in OWR

ARROW

feedback

RRO looks for rightholders

RRO validates the rights status of the cluster

All licensing performed business related legal framework are according action

RRO matches the request against own repertoire (if applicable)

5

License issued by RRO

Request redirected to author/publisher

6

4

Request returned to the library for direct contact to publisher/authors

RRO

ARROW

check authors authority files and adds information about the authors including date of death.

BOOKS IN PRINT

BIP retrieves in print/out of print status, new books and publisher information

ARROW

VIAF

TEL performs a 1st clustering of the records

feedback

TEL CENTRAL INDEX

TEL matches the request with the TEL central Index

feedback

ARROW

LIBRARY

RRO Domain Requirements/1

☐ **That an RRO or similar collective management organisation exists covering the territory concerned**

- ◻ **STATUS:** essential

- ◻ **REASON:** to retrieve information on rightholders (mandating/non mandating, known/unknown, traceable/untraceable)

☐ **That the RRO system is capable of ingesting and outputting standard ARROW system messages using standard communication protocols**

- ◻ **STATUS:** essential

- ◻ **REASON**: to allow machine to machine interaction

RRO Domain Requirements/2

□ **That the RRO has a reliable and comprehensive repertoire database**

  ◘ **STATUS:** highly desirable

  ◘ **REASON**: to support the search process automatically on the basis of the information provided in ARROW messages

□ **That mechanisms exist for the licensing of the digitisation of whole works – either through collectives or through direct licensing by rights holders – for orphan and out-of-commerce works**

  ◘ **STATUS:** highly desirable [but dependent on legislative solutions and rightsholder agreement]

  ◘ **REASON**: to complete the search process with a licensing process

Integrate an existing RRO/1 General Approach

- RRO has to:
  - Provide RRO Query Web Service (M7Q)
    - Set up the B2B interface in order to receive M7Q
    - Test B2B interface
  - Provide RRO Response ARROW Web Service Client
    - Implement/Test Web Service client in order to send to ARROW system the query results (M7R)
  - Design/Implement business logic
    - Elaborate ARROW M7Q
    - Send results to ARROW WS
    - Retry mechanisms

- CINECA has to:
  - Test RRO Query Web Service
  - Implement/Test mapping of RRO query output metadata format in ARROW SH
  - Integrate new Connector in ARROW Workflow
  - Test whole workflow
  - Provide support to RRO during development/test phases

Integrate an existing RRO/2
RRO Specific Approach

- RRO system must anyway be able to receive and elaborate (implement business logic) M7Q and output M7R

- In case just exchange protocol (different from SOAP)/modality (asynch/synch) vary, CINECA must perform ad hoc analysis/implementation

Integrate an existing RRO/3

- What does the RRO need to do
  - Need to have the information available: repository, manifestations and rightholders. Either in a decentralized environment or in a central db as arrow seems to be proposing
  - Need to have mandates pre-recollected, ideally in the same system, to assist the decision of capability to licence requested manifestation
  - Good knowledge of legal framework, particularly in the area of out of print and orphan.
  - A clear catalogue of services provided by the RRO. Licence, find rightholders, perform DS?

Integrate an existing RRO/4

- What does the RRO need to do
  - Study Arrow documentation and understand the complete workflow. Very important, get the common knowledge of nomenclature and semantics.
  - Understand what arrow expects from the RRO and what is left out.
  - Define the internal decision tree: how decisions regarding status and licensing are to be taken, what can be automatised and what must be kept manual.
  - Does the RRO hold all the information needed to reply to Arrow queries?

Integrate an existing visual artists' CMO/1

WP4/6

- Prerequisites
  - Assess the compliance of ARROW metadata schema with the Visual Artists' CMO requirements and in case make the necessary modifications
  - Impact of Visual Artists' CMO in ARROW workflow and it's relation with the RRO of the same country to be established
- CINECA has to:
  - Set up a Web Service Client for sending requests to Visual Artists' CMO
  - Set up an asynchronous Web Service to enable Visual Artists' CMO to send elaboration output
  - Implement/test metadata format mapping if necessary
  - Analyse the impact of this new category of data provider on the ARROW workflow
  - Review the business logic accordingly
  - Integrate new Connector in workflow
  - Test whole workflow

Integrate an existing visual artists' CMO/2

WP4/6

- Visual Artists' CMO has to:
  - Provide Asynch Web Service to receive ARROW messages (queries).
  - Analyse/Implement/Test business logic for query elaboration
  - Provide Web Service client to send query results to the ARROW Web Service
  - **NOTE:** in case the Visual Artists' CMO has to be interfaced with other RROs of the same country, the scenario gets more complex and requires detailed analysis

Set up a new RRO

- What needs to be done
  - Set up of RRO system
    - Analyse RRO system requirements (CINECA/CEDRO)
    - Design RRO system architecture (CINECA/CEDRO)
    - Implementation/Test of RRO System
  - RRO feeding
    - Analyses of existing resources
    - Realization of an import tool from existing source
    - Data mapping
  - Analysis of RRO integration in ARROW Workflow

When the complete workflow is in place

- What needs to be done
  - Choose the validation model (simple, default or advanced) for each country/workflow
  - Coordinate the selection of records and the creation of ground truth
  - Organise usability check including interviews
  - Evaluate result, provide feedback to technical development, write reports

- What do partners in the country need to do
  - Interact with national contact points
  - Select representative records as input for validation
  - Contribute to setting up ground truth data
  - Take part in usability tests

ARROW

FURTHER INFORMATION

PAOLA MAZZUCCHI

Associazione Italiana Editori
(Project Coordinator)
Corso di Porta Romana 108
20122 Milano
Italy

Tel +39 02 89280829
Fax +39 02 89280863

paola.mazzucchi@aie.it
Skype: AIE_paolam
Twitter: @MetalGoddess

ARROW

http://www.arrow-net.eu

ICTPSP
ICT POLICY SUPPORT PROGRAMME