Aggregation Strategy

(MS68 Metis strategic recommendations M18)

Revision	1.1
Date of submission	20 May 2020
Author(s)	Andy Neale, Europeana Foundation; Valentine Charles, Europeana Foundation
Dissemination Level	Public

Revision History

Revision No.	Date	Author	Organisation	Description
0.1	16-12-19	Andy Neale, Valentine Charles	Europeana Foundation	Draft
0.2	04-05-20	Andy Neale	Europeana Foundation	Updates
0.3	19-05-20	Antoine Isaac, Henning Scholz, Julia Fallon, Albin Larsson, Harry Verwayen	Europeana Foundation	Review
1.0	20-05-20	Andy Neale	Europeana Foundation	Final version for DCHE Subgroup feedback
1.1	09-09-20	Andy Neale	Europeana Foundation	Updated to reflect discussion with DCHE Subgroup

TABLE OF CONTENTS

Executive summary

This strategy aims to provide medium-term direction for the aggregation of European cultural heritage metadata and content.

In addition to supporting the needs of the Europeana APIs and website, technical developments for Metis also need to consider the digital transformation needs of CHIs and Aggregators inline with the Europeana Strategy 2020-2025.

Aggregators need to see faster and more efficient publishing options, and CHIs need to see more value from being involved.

The expectations of increased speed, improved data quality, and continual growth will put significant pressure on services.

Based on user research six use cases were explored, covering the use of current services, more regular updates, faster publishing, easier onboarding, improved data quality, and improved content support.

In response to these use cases, seven goal areas have been proposed as an interlocking solution to focus on over the coming years:

1. Maintain the current Metis service
2. Speed up dataset updates
3. Involve contributors in testing
4. Enable fast track publishing workflow
5. Add new data source options
6. Encourage data enrichment
7. Investigate content hosting

Central to this strategy is the concept of an *Extended Sandbox*, that will provide Aggregators and CHIs with ways to speed up the publishing process, and support digital transformation with new reporting and enrichment tools.

The roadmap identifies logical groupings and sequences of work to demonstrate that implementation of the strategy is achievable, subject to prioritisation and resources. Significant progress is expected over a two year period.

Introduction

Purpose of this document

The purpose of this document is to provide medium-term direction for the aggregation of European cultural heritage metadata and content. This is to ultimately support the mission of the Europeana Initiative, but also to provide practical guidance for Metis, the current Europeana aggregation platform.

Background

Metis[1] is the platform used by the Europeana Foundation to manage the aggregation of collection datasets from Aggregators and Cultural Heritage Institutions (CHI). These datasets are made up of metadata records, with corresponding references to digital representations of content objects. The first version was developed to satisfy requirements for the publishing of aggregated collections, via Europeana APIs, to the Europeana website.

The Metis platform was launched in November 2018. It performs data import, validation, normalisation and enrichment, technical metadata extraction and data indexing in a workflow based-system to ensure uninterrupted data publication.

The current operating model mainly supports the delivery of data from domain, thematic and national aggregators to Europeana. Aggregators act as the main contact point with CHIs and perform data operations on their behalf. Metis is used by the Europeana Foundation to then process the data from Aggregators, before it’s published to APIs and europenana.eu. Both Aggregators and CHIs are seen as contributors to Europeana.

[1] https://metis.europeana.eu/dashboard

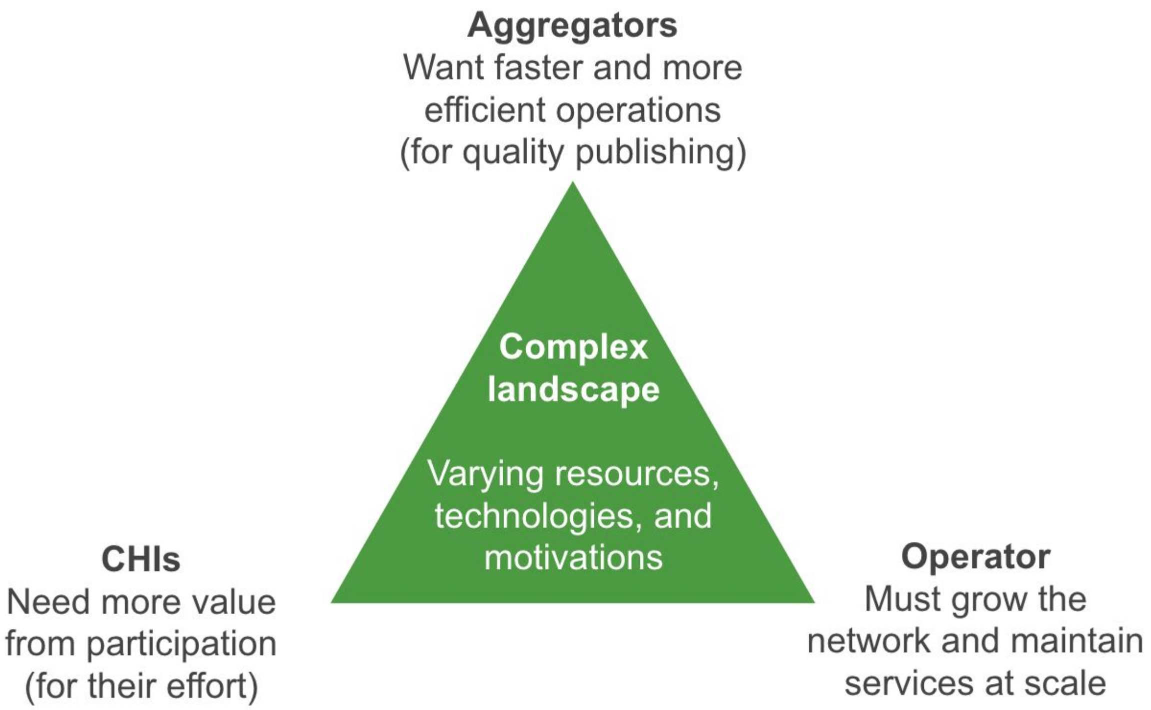
In line with the Europeana Strategy 2020-2025[2], future technical developments of the Metis platform will also need to consider the digital transformation needs of CHIs and Aggregators. This is in addition to the evolving needs of the Europeana APIs and website.

This strategy document starts from the viewpoint that Metis is currently a platform to support aggregation of metadata and content. The cultural heritage sector has different needs for data publishing that may go beyond the requirements of a specific tool like Metis. It is for this reason that this strategy looks more broadly at aggregation and its contribution as a whole.

Problem space

The landscape for aggregation is very complex, with contributors all having varying resources, technologies, and motivations. Of particular note is that CHIs need to see more value from being involved in Europeana if they are to be motivated to take action. Particularly as it relates to improving data quality.

While Aggregators are also concerned with value, it's the operations of Europeana that have a big impact on their work. The need for faster and more efficient publishing being paramount to them. As the operator, the Europeana Foundation must also try and build new services, while maintaining, and growing its offer.


Aggregators
Want faster and more efficient operations (for quality publishing)
Complex landscape
Varying resources, technologies, and motivations
CHIs
Need more value from participation (for their effort)
Operator
Must grow the network and maintain services at scale

[2] https://pro.europeana.eu/post/europeana-strategy-2020-2025-empowering-digital-change

Strategic drivers

Given the complex landscape, it is important to be clear about the factors that are strongly influencing the direction. The factors driving this proposed strategy are:

1. Contributors expect an easier and faster publishing process
2. Contributors need more help improving the quality of their data
3. Growth in the quantity of content types, collection size, and contributors is expected
4. Focus on quality over quantity is to be maintained
5. The current operating model will not scale well further
6. Future technologies are expected to disrupt the current approach at some point
7. CHIs will continue their digital transformation based on their capabilities and resources

User research

Several research actions were taken to provide input into the development of this strategy:

1. Review of previous research from Europeana
2. Review of Aggregator landscape from Common Culture project
3. Interviews with cross-section of Aggregators and CHIs
4. Validation with key stakeholders

Not all research considered is able to be published. The Common Culture project for example has not published its findings yet, but an early review has informed this strategy. Also see Appendix A for a summary of interview notes with Aggregators and CHIs.

Use cases

After considering the user research outputs, and the practical needs for aggregation in the medium-term, seven main use cases were identified. These use cases propose the best areas to focus on for driving the aggregation strategy. It has been found that CHIs and Aggregators want to:

I. **Use current services**

Aggregators currently collect and curate digital collections from their CHI contributors according to Europeana requirements. They then work with the Europeana Foundation to publish the aggregated collections to the Europeana website via Metis. This is an operating model that currently works for many aggregators and CHIs, albeit with the expectation of service improvements over time.

II. **Update datasets more regularly**

Some aggregators receive regular updates to existing published datasets from their CHI contributors. Processing currently requires the entire data set to be ingested again, even if the ingest settings are the same, or only a few fields are added or changed. There is an expectation of easier updates in this case.

III. **Publish more quickly**

Aggregators and CHIs ultimately want to see material published to the Europeana APIs and website more quickly. There is often a bottleneck in the Metis testing and workflow process, where any identified issues require a back-and-forth communication process that can significantly delay publishing. Measures are needed to unblock or automate the workflow process.

IV. Have an easier onboarding experience

CHIs and Aggregators need better ways to transform and share their collections with Europeana, preferably without the need for additional technical infrastructure. There is currently a significant barrier to entry for many CHIs needing to conform to Europeana or Aggregator requirements. It can take a lot of effort that may not be perceived as valuable enough to break through a capability or resourcing barrier. Metis is at the present time an internal tool for the Europeana Foundation, and in and of itself does not necessarily make the process any easier for CHIs who currently work with Aggregator infrastructures. Easier experiences for joining Europeana are needed, while also maintaining more sophisticated options for established contributors.

V. Improve data quality

Seen through the lens of the Europeana Publishing Framework and the evaluation of metadata and content quality tiers, there are gaps in the quality of contributed data across the aggregated collections. But beyond the simple tier measures, there is also the recognised value of enriched metadata to support improved discovery methods, contextual understanding, and data linking opportunities. There is a need for more tools and processes to complement the work that CHIs and Aggregators are already doing.

VI. Improved content support

Aggregation is an activity that not only includes metadata, but also the media objects that go with them. There is an increasing need to support the changing needs for accessing and interacting with images, audio visual items, 3D, and full text that should be considered in the context of aggregation.

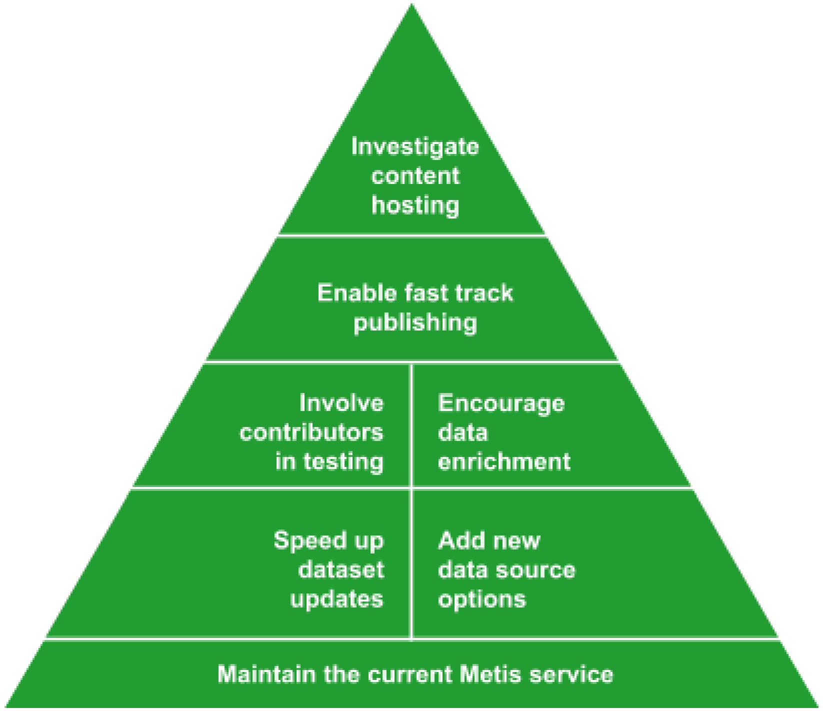
Conceptual solution

Approach

The starting point for the proposed solution is the understanding that Europeana needs to not only meet the needs of Aggregators, but also support growth of CHI capability inline with the Europeana Strategy 2020 - 2025. To this end seven outcomes have been identified in response to use cases:

1. Maintain the current Metis service
2. Speed up dataset updates
3. Support contributor testing and preview workflows
4. Enable fast track publishing workflow
5. Add new data source options
6. Encourage data enrichment
7. Investigate content hosting

The full conceptual model is made up of interlocking solutions for each outcome area, ultimately demonstrating how aggregation services might evolve over the coming years.

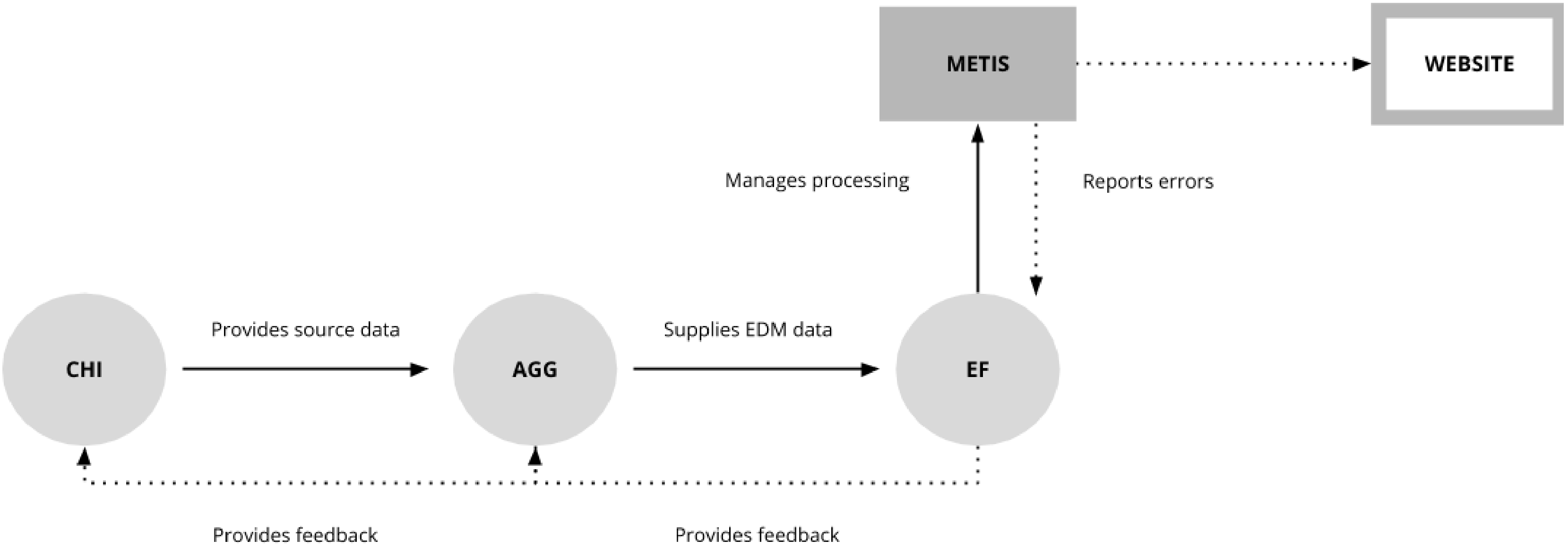

Investigate content hosting
Enable fast track publishing
Involve contributors in testing
Encourage data enrichment
Speed up dataset updates
Add new data source options
Maintain the current Metis service

1. Maintain the current Metis service

The current operating model for aggregation works for many contributors, and should be maintained to keep services running as expected. Under the current model, Aggregators act as an intermediary between CHIs and the Europeana Foundation, and the Metis Platform is used by the Europeana Foundation to ingest EDM records for processing for publishing.

In this strategy it is proposed that Aggregators have the option to keep using this model, while also being able to take advantage of new features described in the full conceptual solution outlined in this strategy. Alongside the current operating model, new ways of publishing to Europeana are proposed in the following sections that Aggregators may choose to adopt if they wish. This largely centres around the ability of CHIs to use the Metis Platform directly to support their own objectives. It is expected that most CHIs will still want or need the support of an Aggregator, but that some CHIs may choose to publish directly with Europeana.

Current operating model

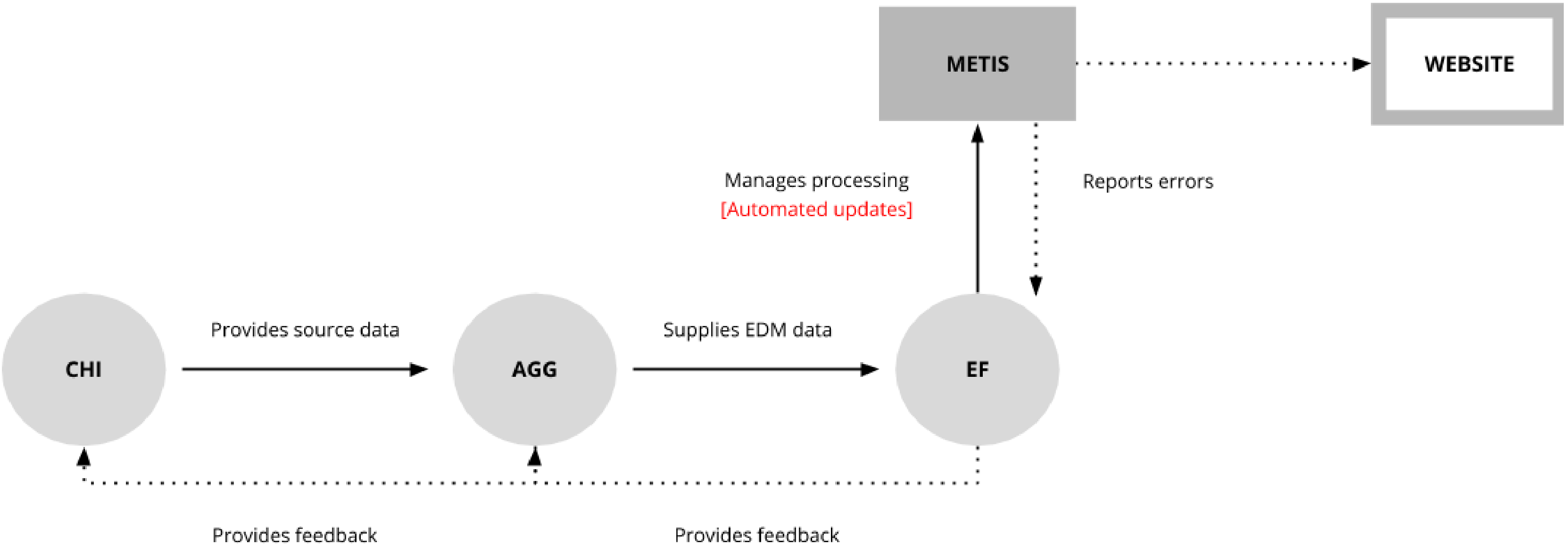


2. Speed up dataset updates

To process dataset updates the current Metis Platform needs the entire dataset to be ingested again, even if the ingest settings are the same, or only a few fields are added or changed. This solution component will see publishing speed increase by allowing the processing of only changes to the dataset. This is called incremental ingest, and then allows the establishment of scheduling, which will allow datasets to be published to the Europeana website on a regular basis with limited human intervention.

Incremental ingest requires Aggregators or CHIs to establish publishing systems that can automatically provide updates to Metis, through the use of APIs or protocols such as OAI-PMH. Ultimately this may be implemented by CMS vendors. Scheduling requires that an existing dataset workflow has been established, and that scheduled updates successfully pass validation and workflow steps.

Automated dataset updates



Outcome	Speed up dataset updates	
Users	Aggregators (with possibility to open up to CHIs)	
Main features	Incremental updates	Scheduled updates
	● Accept the update of records for an existing dataset ● Supports the addition, deletion, and update of individual records within a dataset ● Does not require the reprocessing of the entire historical dataset	● Can reprocess a dataset based on a scheduled workflow process ● Workflow process can run automatically and be set to publish to live environment without human intervention ● Schedules can be established on a recurring basis. Starting at a particular time/day and recurring every *x* hour/day/week/month
Longer term potential	● Component of further automation process	● Component of further automation process

3. Support contributor testing and preview workflow

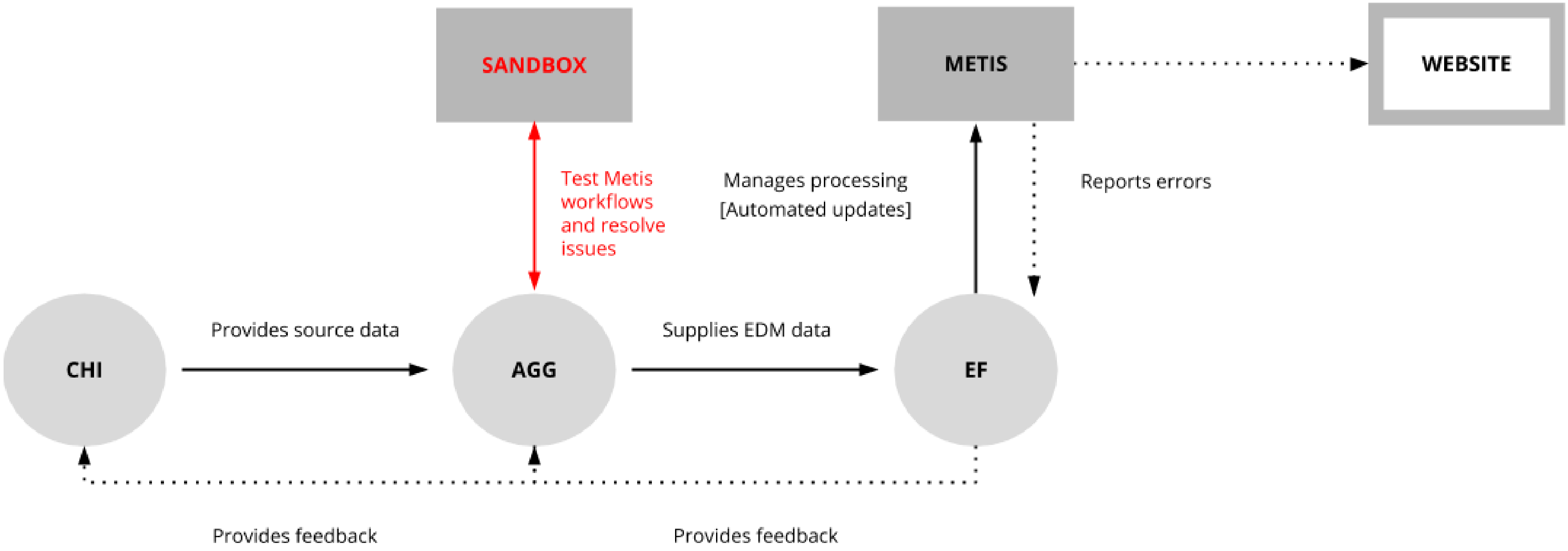
This strategy proposes introducing a *Sandbox* environment to allow Aggregators and CHIs to participate in testing and preview workflows to resolve publishing issues before processing by the Europeana DPS team.

Users of the *Sandbox* would be able to trigger the workflow from data import all the way through to previewing in a Europeana website environment. It would make the processes such as data import, validation, normalisation, enrichment, and preview, fully transparent.

Aggregators and CHIs could therefore make sure their data is compliant with the Europeana requirements, and can immediately resolve any issues related to EDM data validation, missing media, or broken links for example. It would be expected to decrease the amount of time taken with back and forth communication currently taking place between CHIs, Aggregators, and Europeana.

In order to support this effort users of the Sandbox would need access to a wide range of online guides, tutorials, webinars, and error reports.

Self-testing of workflows

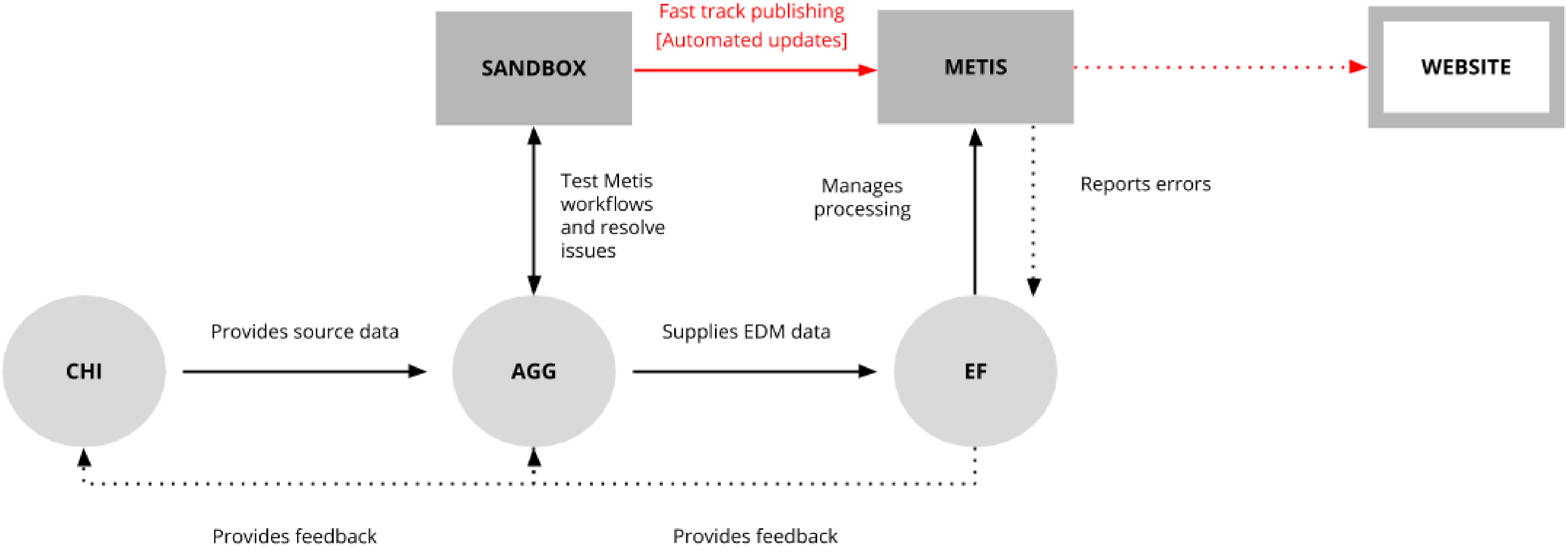


Outcome	Involve contributors in testing	
Users	Aggregators (with Agregators sharing with CHIs if they wish)	
Main features	Test of full publishing workflow	Europeana website preview environment
	● Execution of Metis workflow including data validation, normalisation, enrichment, media service and indexing in the same condition as Metis	● Inspect the results of data and media processing before submitting final publication request
Longer term potential	● Integration with Metis for fast-track publishing	

4. Enable fast track publishing workflow

By moving to incremental and scheduled updates, and on the basis that errors are corrected in the *Sandbox* before being submitted for publishing, a more automated workflow should be expected. In practice it means that all the steps of data publication will be triggered in one workflow and Europeana Foundation staff will only do a final acceptance check when entirely new datasets are being added. Updates for datasets where the configuration hasn't changed should need no staff review at all. Notifications can be put in place to ensure that Aggregators and CHIs are automatically informed about the status of their publications. With further integration between Metis and the *Extended Sandbox*, a fast track publication process can be enabled.

Reduce human intervention



Outcome	Enable fast track publishing workflow
Users	Aggregators (with Agregators sharing with CHIs if they wish)
Main features	Metis automation
	● One single workflow with final acceptance at the end of the process ● Publishing to Europeana website can be triggered within Sandbox ● Automatic notifications and publication status updates

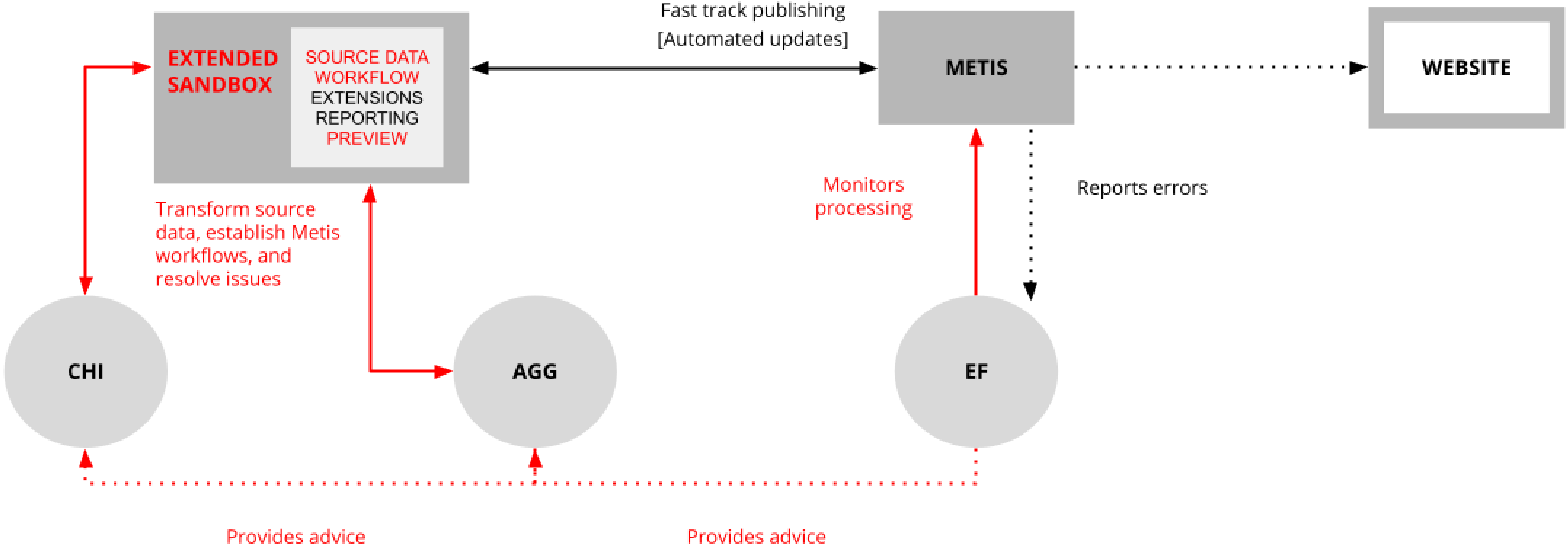
5. Add new data source options

Once the basic *Sandbox* environment is established, it can then evolve into a full pre-publish environment where CHIs can work with or without Aggregators to transform source data, establish Metis workflows, resolve issues, and fast-track publishing. This would take the tool beyond a simple sandbox, so the working title of *Extended* Sandbox is used here to make that point.

New features such as uploading data and transformation from common standards would be established to give the Europeana Initiative an additional way to support CHIs in their digital transformation. The intention is to both lower the barrier to entry for smaller CHIs, as well as provide additional transformation options for aggregators that can speed up their processes.

The *Extended Sandbox* would no longer only expect the Europeana Data Model (EDM) as a data source input. It would add new data source options to support the conversion to EDM directly, or ingest EDM metadata via other data exchange technology. To make sure the solution can scale over time, the new data source options would be limited to specific data standards more commonly used by contributors, and standard transformation rules would be applied without support for XSLT editing. The ability to upload custom XSLT files, and integrations with other more sophisticated tools, like MINT, would still ensure a full range of options for contributors. See Appendix B for a conceptual model of the *Extended Sandbox*.

Data source options



If CHIs do start using the *Extended Sandbox* to publish more data through to Europeana, then further changes to the support model and infrastructure will be needed. The current operating model for aggregation relies on the Europeana Foundation having direct support relationships with each Aggregator. That direct relationship for support of Aggregators should continue, however it doesn't scale if direct support for CHIs using the *Extended Sandbox* is also required. Instead, the use of a community forum, where CHIs can ask peers for advice is proposed. The establishment of a virtual ingest team could also be considered, made up of Europeana Foundation and Aggregator ingest specialists who can monitor and support the community forum.

Outcome	Add new data source options	
Users	Aggregators and CHIs	
Main features	Source data import	Transformation options
	● Upload CSV file (must use template) ● Upload XML file ● Harvest data from MINT, Linked Data or IIIF source	● CSV template designed to map easy to understand field names directly to EDM ● Default XML transformations proposed for DC, MARC, LIDO, and EAD ● Ability to upload custom XSLT file for transformation ● No graphical mapping tool (MINT available for that purpose)
Longer term potential	● Support for more data import choices can evolve over time depending on need	● Support for more transformation choices can evolve over time depending on need ● Develop a data storage system (including versioning and archiving)
Support method	Guides, tutorials, webinars, community forum	

6. Encourage metadata enrichment

Using the *Extended Sandbox* to make it easier and faster for publishing to Europeana is an important step, but data quality goals that have driven recent efforts must also be addressed. This strategy proposes two areas to progress in the context of aggregation. Firstly, to improve the availability and comprehension of reporting data, so that CHIs and Aggregators know what to act on. Secondly, to harness the power of artificial intelligence and machine learning to enrich data.

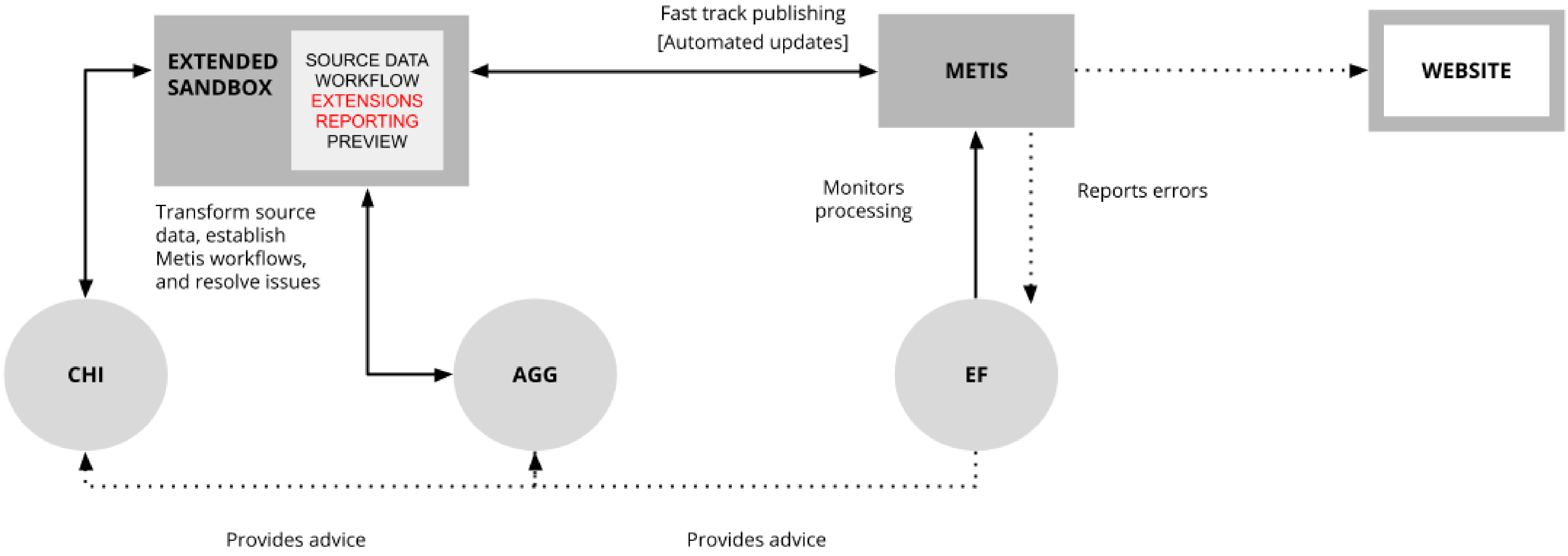
Through the Europeana Publishing Framework, Europeana has developed a tier system for metadata and content that aims to help Aggregators and CHIs decide where to focus their data quality efforts on. Contributors are now in need of more statistics and error reports that can help them in their decision making. This would be expected to include further insights into their tier statistics, and additional error reports from validation and media service processing similar to the ones produced currently by Metis. But perhaps more importantly, reporting should include opportunities to increase discovery of items, with more insights into the long tail of data issues[3] such as broken links, normalisation issues, and rights coverage for example.

Encouraging data quality improvements also means providing Aggregators and CHIs with the means to act on the insights they develop. With the rise of artificial intelligence and machine learning capability over the past years, it has become clear that data quality improvements no longer need to rely solely on cataloguing practice within CHIs. The next logical step for the *Extended Sandbox* service is therefore to include more enrichment and validation processes in workflows. This would be in support of both improving Europeana's publishing outcomes, but also as support for digital transformation if CHIs wanted to use these processes to improve data in their own Collection Management Systems.

Metis does already allow for this to some extent, but the intention would be to add an extension module that CHIs and Aggregators could use directly. Enrichment and validation would be first tested in the *Extended Sandbox* workflow to determine whether they should be included in the production workflow, and then executed as appropriate. Integration with enrichments and validations that occur outside of Metis, before ingestion, also need to be factored in. Plugins for features such as language detection, entity tagging, image detection, and automatic transcription for example should be made available over time. See Appendix B for a conceptual model of the *Extended Sandbox*.

[3] Longtail of data issues report available on request (MS74)

Additional enrichment and reporting tools



Outcome	Encourage metadata enrichment		
Users	Aggregators and CHIs		
Main features	Errors reports during workflow executions	Content and metadata reporting	Enrichment extensions

	● Reporting of Sandbox workflow including data validation, normalisation, enrichment, media service and indexing in the same condition as Metis	● Insights into tier statistics ● Data quality opportunities to increase discovery of items ● Long tail of data issues such as broken links, normalisation issues, and rights coverage	● A library of enrichment extensions would be available to choose from ● Options would evolve over time, but could include language detection, entity tagging, and image detection ● Once the results have been considered the extension can be turned on/off during the Metis workflow. ● Integrate pre-processing enrichment workflows from Aggregators ● Provide support for CHIs to extract enriched data
Longer term potential	● Integration with Metis for fast-track publishing	● The reporting of data issues and statistics can evolve over time.	● R&D across the Europeana Network could contribute to these extensions ● Could allow better integration of appropriate Generic Service developments to occur

7. Investigate content hosting

The Metis Platform itself is not a hosting environment. It instead processes both metadata and content and makes them available to different systems, such as search indexes or thumbnail repositories. While the delivery of content objects (sometimes referred to as *media*) is not a function of Metis, it is a fundamental need for end users that should be considered in the context of aggregation.

This strategy takes a medium-term view that the cultural heritage sector might benefit from having additional content hosting options so as to increase the accessibility of digital content. From the Europeana perspective this would be about lowering the barriers of entry for smaller CHI, but also to support the digital transformation of the sector as a whole.

Other Europeana systems currently host some full text content, such as scanned images of newspapers, and their OCR text equivalents. As well as transcriptions of different types. These however have been developed to support one-off projects and are not implemented to support ingestion and hosting at scale. It has also been suggested that the cultural heritage sector needs shared environments for the likes of IIIF image hosting, audio/visual media streaming, and 3D object hosting as it is not always practical for smaller institutions to establish their own environments. In these cases, Europeana should investigate the options for providing these facilities to CHIs.

Outcome	Investigate content hosting	
Users	Aggregators and CHIs	
Main features	Full text ingestion	Media hosting
	● Investigate newspaper and full text content hosting	● Investigate IIIF image hosting ● Investigate media streaming hosting ● Investigate 3D object hosting

Roadmap

The roadmap starts to identify, organise, and sequence the types of tasks that are required to deliver on this aggregation strategy. Scheduling of activity will happen in implementation planning, subject to prioritisation and resources. Significant progress would be expected over a two-year period. Significant communication and consultation with contributors and stakeholders would need to occur throughout the implementation process.

Outcomes	**Stage 1**	**Stage 2**	**Stage 3**
Speed up dataset updates	Metis is updated to support incremental updates from OAI-PMH endpoints. Develop case study to demonstrate the cost/benefit value to providers who might upgrade their technology to support incremental updates. Revise Metis user interface to optimise the approval process for updates Design specification of ingest API to support incremental updates supporting the addition, deletion, and update of individual records within a dataset.	Scheduling features allow dataset updates to be established on a *one off basis.* Notifications and status updates are provided to the users. Validate ingest API with proof-of-concept (experiment).	Scheduling features allow dataset updates to be established on a *recurring basis.* Starting at a particular time/day and recurring every x hour/day/week/month. Ingest API is enabled for Metis. Integration guidelines are developed for systems that want to take advantage of incremental updates.

Involve contributors in testing	Sandbox APIs support a full Metis workflow including data validation, normalisation, enrichment, media service and indexing under the same condition as Metis. Sandbox API includes reporting for data validation, normalisation, enrichment, media service and indexing under the same condition as Metis. A user can preview the published data on a copy of the Europeana website	Sandbox website is built and integrated with Sandbox APIs Sandbox user guide is developed. User testing of Sandbox performed as input into Extended Sandbox planning	
Enable fast track publishing workflow		Validate that incremental updates can be published to the live environment without human intervention (experiment). Implement user login and account management features. Notifications and status updates are extended.	Existing Metis services are optimised to avoid unnecessary processing (for instance during a dataset update a thumbnail may not need to be generated again if it already exists). The Sandbox and Metis are better integrated to create a fast track publication route. A user can specify the settings for the Metis processes from

			the Sandbox.
Add new data source options	Validate the modelling of CSV upload process (experiment). Validate the transformation of one select source data formats to EDM via pre-set XSLT mappings (experiment). Implement CSV and select data source transformations in workflow. Purge uploaded source data on a regular basis.	Implement select data source transformations in workflow. Support upload of custom XSLT mapping for transformations. Enable workflows to acquire data and processing rules from MINT.	Enable joint storage layer for the Sandbox and Metis to optimise infrastructure and add support for archiving and versioning. Validate the need for support of additional data imports via the likes of IIIF, Schema.org, or linked data sources.
Encourage data enrichment	Tier statistics are built directly into the Sandbox so users can calculate their tiers before publishing the data in Europeana.	Data quality reports extended to include longtail of data issues such as broken links, normalisation issues, and rights coverage. Design the extension architecture so that additional enrichment and validation processes can be connected with the Sandbox. Propose updated Europeana Website designs and EDM	Design and test the user interface for selecting, testing, and using extensions for enrichment. Evaluate enrichments from third-parties and generic services for possible integration. Implement workflows that allow users to test and trigger enrichments.

		changes that take account of further enriched data. Validate the extension architecture with a proof-of-concept for one enrichment process.	Develop extension integration guide.
Investigate content hosting	Investigate newspaper and full text content hosting	Investigate 3D object hosting	Investigate IIIF image hosting Investigate media streaming hosting
Infrastructure	Architecture and infrastructure planning to support future Metis needs.	Performance upgrades to support growth of Sandbox and Metis usage.	

Appendices

Appendix A: User research interview notes

These are the rough notes from interviews undertaken in December 2019 as input into this strategy. A cross-section of Aggregators and CHIs were interviewed.

1. Make it easier to *prepare* data for Europeana

- CHIs need training to be able to understand crucial data details.
- CHIs are usually not familiar with technologies and this is why so far tools like MINT have brought a simple way to define mapping.
- CHIs also have to understand how to change data in the Europeana Data Model.
- For Aggregators the main task consists of validating data for later publication in Europeana and then bringing back the needed changes to data providers. It is therefore important for Aggregators to have more access to preview environments. Allowing them to be more independent in the checking of the data.
- If Aggregators could do more of the data checking on their own it would add some extra motivation. By data checking here I mean a data check similar to the one Europeana runs in addition to the data quality control routine in place on the Aggregator side.
- CHIs need to understand the standards which are at the moment the main issues: Europeana Data Model, RightsStatements but also EPF.
- Aggregators need to make sure data requirements are applied to the correct use case. It can help to promote collaboration between Aggregaotrs or CHIs and encourage the reuse of content and share use cases so that they can better identify actions on the data.
- CHIs care about what happens to the data and it is why CCO license is still an issue.
- The interest in the data is different depending on the professional liaising with the aggregators. Some are more interested in the display of the data or integrity of the data.

- Data improvements needed by Europeana are understood by CHIs but they might not be in the position to apply changes
- CHIs are happy to review their data but the effort they put in it should be demonstrable. It is not the case at the moment: your content is still lost among other materials.
- Use data providers to nudge each other and serve as examples. CHIs are usually more relaxed after the first contribution to Europeana
- Aggregators should provide more insight on a domain as they bring expertise.
- Aggregators have a role of mediation: scope for more training, one to one discussion
- The statistics dashboard provides a good visual representation, a clear idea of the data you can find and how to use it.
    - It would be good to have a link to real examples so we can identify the problem directly.
    - It would be more used by aggregators.
    - What is missing is the help to better understand the tiers and some help to read/interpret them.
    - Listing of values, being able to identify missing elements
    - Zoom in on tier 0; what to do with it, how can we analyse.
    - More breakdown would be useful to see what can be improved.
    - Not everything needs to be solved before publication, important to include the data quality as part of the workflow without necessary positioning it as a blocker.

2. Make it easier to *deliver* data to Europeana

- The most difficult for CHIs is the creation of EDM as CHIs do not have the required technical expertise.
- It would be interesting to facilitate data validation and conversion without having them to execute the EDM schema.
- In the last 10 years there has been very little adoption of OAI-PMH. There are other mechanisms for transferring data that can be explored such as using a browser.
- Advanced CHIs are reluctant to implement a system for Europeana. They want to reuse what they have or what they have paid for (proprietary system).
- The current approach is not scalable and Europeana needs to support additional methods, we need to be more flexible in the approach.
- Whatever approach is chosen it shouldn't be high requirement for providers but also for Europeana
- Technology providers haven't built OAI-PMH as part of their products. Academic libraries are probably the only one who did.

    - Implementation of OAI is not easy for developers because technologies have changed. OAI is XML based and not a RESTful API and not applied on WebResources.
    - Repox was trying in a way to build the REST layer.
- Technologies like IIIF or Linked Data still require efforts from CHIs but bring advantages to them. It is a more win/win situation
    - These technologies can help providers to adapt to the digital age but it is currently not really as a data aggregation alternative mechanism.
    - Schema.org harvesting (internet search engine) can be used to promote the usage of linked data and change the way CHIs work.
- It can be interesting for Europeana to look for technologies that can bring more without focusing on data delivery to Europeana. A similar approach should be taken for data quality.
- Validation and tier calculation is something to investigate.
- Suggest taking the discoverability by Search Engines as starting point (Schema.org and RDFa data pages). The motivation is based on assumptions that the technology is worth investing into.
    - Need to be accompanied by the development of a model for keeping versin, timestamp (DCAT).
    - The idea is that the data should be reusable for different purpose and nothing additional should be required for Europeana by the CHIs. Additional work should be done by Aggregators and Europeana.
    - One idea is to develop a plugin to fetch the data to be used by aggregators and Europeana.
- Issue at the moment is that the data is not very available: there are no practice in defining a dataset and best practices for data aggregation. We need to better describe things.
- In this sense, we need to go back to features of collection management system.
    - Lobbying to publish linked data and embedded URI in object description at source.
    - Term network=use of vocabulary
    - More post processing approach (pragmatic approach)
    - There are 2 components: create data/publication data: it is not clear if it needs to be merged into 2 components or 1 component.
- Using URIs still requires standardisation or URIs. Could work with the current landscape, Still need support for small institutions. Put it in Cloud service→ this is more a role for the Network or Europeana.
- Enrichment can be looked at as a separate service: CHIs could decide if they are interested in those data.
    - Annotation are maintained separately but could be in for sustainable reasons.
    - Quality=completeness for us in terms of replacing literals by entities.

    - We still want to promote published vocabularies and standards. Alignment tools are needed.

3. Increase the speed of data publication in Europeana

- Aggregation landscape ⇒ vision
- Publishing platform→ immediate publishing should be focused
- Expectations are high, use Jira tickets, less manual steps.
- Comes with the control of the pipeline
    - We have to be clear with what we want to achieve: a curated platform (=more portal)??
    - Vision and mission has an influence
    - Visibility where is the data and how it is used in the process→ links to Europeana/errors/
- Important to access the data where they are→ the only role of CHIs is to unlock the data for aggregators.
- Data delivery
    - Integration is costly so development of APIs is a better approach but there could be issue of scalability and sustainability but overall the costs to develop an interface should go down.
    - Proof is in the use. There is a role for aggregators to demonstrate the potential of these services.
    - The prioritisation on what should be integrated should come in a second stage. The work of the aggregators could be used as a base for a bigger market study.
- Data preparation
    - Different engagement with customer
    - IIIF service → not enough to take data. We need to see where the service lives (more requirements.
    - If you deliver a IIIF service CHIs become your users. There are less users for data delivery.
    - When asking ourselves what kind of services should we provide we should think about what people need.
    - We need to promote Cloud services and offer services that would make their current model obsolete.
- We need modular services
    - Infrastructure
    - Content/enrichment/normalisation
    - Content management system/crowdsourcing layers.
- Legacy systems what to do with them?

    - Transition from legacy to innovation → design is part of the product.
- We have to lower threashold→ use quality as a mirror to indicate the severity of the problems and then we need to actions.
- Maybe we should also provide guidelines to have better infrastructure.

4. Motivate aggregators and CHIs to provide higher quality data

- Data quality is framed by a series of frameworks and standards like EPF, EDM. In addition, the DPS team at Europeana will provide reports with the list of errors identified during ingestion.
- If we looked at data quality in terms of 1)raising awareness 2) taking actions
    - What are the tools and services Aggregators and CHIs need to provide better quality data?
    - How the work of the DQC can play a role in the Metis strategy?
- Work of Europeana is to motivate aggregators in order:
    - To give better quality metadata
    - To solve issues in the data.
- Their interest in solving data quality issues can be an indicator about the tiers → are the different sub-components the right ones (language, enabling elements, context), are more subcomponents needed.
- In order to solve issues we need to find more detailed targets such as the problems patterns as worked on by the DQC with features such as having access to records ID of the problematic records, details about the problems, explanations about the problems.
- We should also communicate more incidents to providers with detailed report.
- About the detection it should be early in the ingestion process to prevent wrong ingestion but the action could be post ingestion
- We have to support a bit better the exercise of analysis either in the source data or in the mapping. Could be more motivated for contributors in the issues are looked at source.
- Metadata quality assessment tool is good to complement tiers. More specific view to explain certain details in the data such as fields frequency.
- There are more alignments for the multilinguality components.
- Automatic enrichment we could make it optional and decide when to apply it depending on the data.
- The current Metis workflow could offer more flexibility and granularity
- For instance the enrichment workflow could be more configurable. Maybe a granulary per concept/place /agent/timespan.

- Should full-text and annotation be managed as part of Metis?
- Use of Metis to prepare others data: it could be based on template to support the transformation from different formats to EDM for instance (a MINT for IIIF, an export for wikimedia)
- However we also have to put in place mechanism for harvesting or provide services to people who don't have the technology yet.
- It is not sustainable on the long run to have many APIs to manage.
- We need to understand whether providers are interested in something generic .
- Regarding IIIF, the next request would be a IIIF server to generate a manifest + service to publish content for IIIF server.

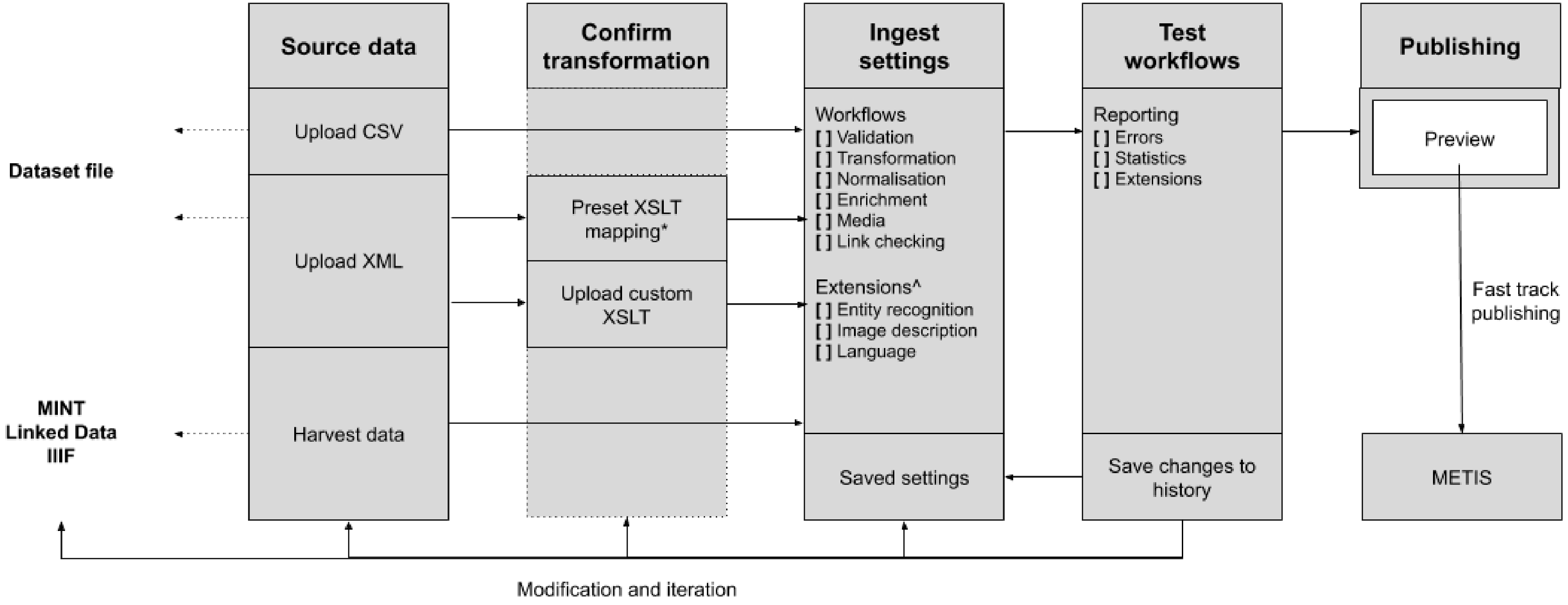
5. Provide a reliable and sustainable infrastructure for aggregators and CHIs.

- LOCAL collections[4] is used by small institutions. There are regional hubs in Poland.
- Core quality comes from CHIs
- As tool provider /aggregator
  - Direct import push to Europeana /API for National Aggregators
  - Mapping: from Common Culture experience we realise we need a tool that simplify mapping activities
- Impact CHI
  - OAI-PMH doesn't need to be costly (basic interface, little data more data is affordable).
- Reason in Poland why there is a digital transformation, is because it is driven by Europeana.
- Small CHIs are already using Cloud based infrastructure. CHIs will not do the transition on their own. They are always used to rent and use services (there is trust and it is affordable).
- Small institutions would use services : issue is financial and sustainability. You need a lot of users to make the service sustainable. LoCloud doesn't have enough users yet.
- IIIF⇒ require storage space and a network with high availability⇒ the main challenge for storage space is not the storage but how responsive he are. PSNC is interested in streaming/ serving IIIF.
- Content storage is of value (IIIF streams):storage at CHIs or PSNC premises .
- Linked data=better to collaborate with content management providers (digital repositories)
  - Source of the problems is in the tool that provide the data.

[4] https://locloudhosting.net/

- Need to manage this data in the system if it is not supported at source. Need to develop alternatives → it requires personnel and skills = training. Need tool in place otherwise it is not needed.
- Enrichment→ integrate tools with some enrichment API (common way to do it)
- Enrichment needs to be done at CHI level → more apis to do enrichment and possibility to apply them only when needed.
- Another service to deliver would be a OAI-PMH programming library→ task to implement OAI harvest. Programming libraries that support OAI exists but are badly supported.
- Overall technical libraries that can be used by bigger number are useful. It is good to integrate communities.
- Overall all technologies require some support. Backward compatibility important. It is easy to follow new technologies for big institutions. More difficult for smaller institutions. New technology is good but we need to think about what is already in place.

Appendix B: Extended Sandbox concept



* Simple mappings from DC, MARC, LIDO, and EAD can be added to over time

^ Plug-ins can include generic service tools and services that improve quality of data

Co-financed by the Connecting Europe
Facility of the European Union

Europeana DSI is co-financed by the European Union's Connecting Europe Facility.

The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.