**1.** Thank you all for the opportunity to speak to you about some work I did with metadata visualization.

**2. Overview**

This is the overview summary of the presentation. I will describe what Artexte and its repository is. I will then introduce the metadata network/graph visualization methodology, and present a series of results.

Then I will conclude with some recommendations for the OR community, based on my experience with this work.

**3. What is e-artexte?**

It is Artexte’s catalogue, and since 2013, repository for contemporary visual arts publications.

Artexte was

* Founded in 1981 as an
* arts organization with the mission to
* Document contemporary art in Canada and facilitate research
* In 2013, the catalogue was transformed into a digital repository based on the EPrints platform.

The collection contains over 28,000 records for contemporary art publications.

In the background of the slide, you see Corina MacDonald’s interface that summarizes the main document types in e-artexte:

* Artist Publications
* Monographs
* Exibition Catalogues
* Thematic Anthologies
* Exhibition Documentation
* Grey Literature
* Periodicals

## 4. Multidisciplinarity of Metadata Visualization

I like the definition of Information visualization as **“***The use of computer-supported, interactive, visual representations of* [*abstract data*](http://www.infovis-wiki.net/index.php?title=Abstract_data) *to amplify* [*cognition*](http://www.infovis-wiki.net/index.php?title=Cognition)*.”* The purpose of it is to assist humans in solving problems. To paraphrase the visual studies scholar Lev Manovitch, “the goal of information visualization is to discover and reveal the structure of a data set that is not known a priori”

The background literature about networks/graphs and their properties is vast and multidisciplinary, including:

* its roots in **mathematics**/graph theory, starting with the work of Euler in the 18th century
* social network statistics and visualization in **sociology** and **sociometrics,** with the work of Moreno and Milgram,
* **scientometrics**, and the study of the Web, with the work of Barabasi.
* **visual studies** and more recently, **data science**

In a presentation like this I obviously don’t have the time to go into the details of the literature, so I move right along to the methodology and tools

**5. Methodology and Tools**

I began with XML exports from EPrints, initially transforming them into graphs with XSLT. However, XSLT is too slow, even for a relatively small dataset like the Artexte collection. I read Russell Jurney’s Agile Data Science at this time and switched to Apache Pig running on Hadoop (framework for distributed storage and processing) largely inspired by reading Jurney’s book . The idea is that I might as well switch towards infrastrcture that is scalable, so that hopefully if I need to expand the dataset, I will not need to switch again. At this stage, since I was interested in visualization primarily… [continued next slide]

**6. Visualization Tools (Slide)**

Instead of opting for application development, I chose to work with existing open source Visualization platforms. I began with Cytoscape, and then switched to Gephi.

These visualization platforms provide a workbench to load data as a graph, to query and filter the data, to introduce layout, to map graph measures to visual properties. The objective was to learn about bibliographic data as graphs, to experiment with layout algorithms, filtering methods, visual mapping of properties, and so on. Importantly, I could also export the results to a web site, either through images or through the Sigma.JS export plugin developed by Oxford/JISC.

I can use these visualization platforms because the data set is just at the limit of what they can handle. In my data set, I stayed well under a million nodes, so this is small to medium data. They would not scale well to “big data” graphs of billions of nodes.

I switched to Hadoop both to speed up and open the door towards scalability, at least on the batch processing step.

**7. Graph Data Formats**

A few words about graph data formats.

I began with Cystoscape’s SIF (Simple Interaction Format) for graph files, which is a variation on the EdgeList format. These are Excel-compatible files with a list of edge definitions, a “source” and “target” node, labelled with an id/string name and a label for the edge/interaction.

When I switched to using Gephi, I also switched to GML (Graph Modeling Language).

GML is a text file format with a JSON-like syntax, widely used and well supported in Gephi and Cytoscape. You see an example of what it looks like on the right, with sections for nodes and edges.

There are a few “standard” network/graph formats like NET, the lingua franca of graphs dating back to the Pajak software, GML, or GraphML that are compatible with visualization platforms, and I hope that repository software can export to these formats in the future.

I will come back to graph formats later. Repositories are moving towards exporting their metadata in RDF format, something that some repositories like EPrints has been able to do form many years; however, working with RDF is difficult and not well supported natively by visualization platforms yet.

**8. Map visualization - publisher locations**

Let’s move on to some visualizations.

Artexte contains publications from art gallery exhibitions from all over the world, including North America, Australia (and Brisbane!), Europe, etc.

Using Apache Pig scripting, I re-shaped the metadata as a CSV file with geotagged locations, and then uploaded to Google Fusion tables which can then be converted to a Google Map. It is an easy way to navigate a map of publisher location.

The publisher locations in e-artexte are useful, because the publishers are actually usually art galleries and art centers, etc., where a particular exhibition took place.

The result is a map interface with over 13 hundred unique locations, with anything between 1 and 4742 item links per location (the largest is Montreal, unsurprisingly, other centers are Paris, Berlin, London, New York).

## 9. Metadata visualizations with e-artexte data - Entire collection as a bi-partite network

I was very curious as to how the metadata looks like as a graph.

I began with attempts to visualize the metadata as a bi-partite network; that seemed like the simplest way to go from a list of metadata records to a visual graph.

In this representation, there are nodes representing items ( I used ID/title to label these uniquely ), let’s call these “item nodes”; and nodes for artist names, author names, publishers, critics, keywords and so on, let’s call these “property nodes”. Item nodes are never connected to each other directly, only through property nodes, when for example 2 items have the same keyword. Similarly, two property nodes are never connected directly to each other, only through item nodes.

Once I had a graph, I extracted statistics about the graph, for example, the path length distribution between nodes. You have probably heard of “six degrees of separation” and like in Milgram’s mail experiment that originated this idea, I confirmed that the user is, **on average**, six clicks away from any particular author name, artist name, keyword, etc. while browsing the catalogue.

## 10. Metadata visualizations with e-artexte data (2) – filter 1

The advantage of bi-partite representation is that it keeps the number of edges relatively low, but this graph is still over a hundred thousand nodes and close to half-million edges.

A filtering process is required, something equivalent to a relevance ranking algorithm. I used the “giant component” filter, which focuses in on the main component. In a metadata graph, there are usually some small disconnected components and one main component, connecting the overwhelming majority of nodes. I also filter based on node centrality, narrowing down the graph to 700 or so nodes and under six thousand edges - which is what you see on the slide. This is still a lot, but we can begin to see some structure emerging.

I mapped the betweenness centrality to the node/label size (the larger the value, the larger the node/label). Betweenness-centrality is a measure of how important the node is to the average distance between all nodes in the network – a representation of how significant a connection that node is in the structure.

Right in the center, labeled in red, you see a cluster with a well known Canadian artist named Michael Snow in the middle of it. Off to the left, there is a community with the predominance of prominent keywords, such as photography, architecture, design, identity, etc. What is useful about bi-partite representations is that they retain all of the items, authors, keywords, etc in the graph, without exponentially increasing the number of edges.

## 11. Metadata visualizations with e-artexte data (3) top 141 nodes

Increasing the centrality filter further we get something that can be processed visually more easily. This graph has 141 most central nodes for keywords, (photography, landscape, installation, architecture, abstraction, etc) and items in the collection, such as the “Tableau Inaugural..” a major exhibition in Montreal in the early 90s to celebrate the opening of the new location of the Museum of Contemporary Art. This graph is one answer to the question, what are the most important pieces of metadata in the collection?

**12. Metadata visualization – Bipartite - Sigma.js export from Gephi**

What you have seen up to now are images generated with Gephi. Sigma is a JavaScript library that makes it possible to publish networks to the Web in an interactive way.

The export was developed by Oxford Internet Institute and JISC, and it allows me to export a Gephi graph to the web, and include a search box, an information box for each node, that includes all of the nodes that it is connected to. In effect, for this example of a bi-partite metadata graph, it results in an experience that is close to navigating a catalogue.

The user can search for and select a node, like Michael Snow, with a panel on the right, listing all the items that the artist is a part of, as well as select entire clusters/communities, such as the 800 node cluster that Michael Snow is a part of. The user can zoom in or out on the network, and the visualization shows the appropriate level of detail at each zoom level.

## 13. Metadata visualization – co-authorship network

Next, I began to experiment with single-mode networks. That means making projections. For example, in collaboration network, if two authors are listed as co-authors on an item, then that is one edge between them. While in bi-partite representations an item is a node, in these projections, the item is represented by an edge.

Co-authorship or collaboration networks are often highly clustered, and we see that in this case as well.

Force-directed layout algorithms model the graph as a physical system, with forces for attraction and repulsion. The layout that I used for most of these is Gephi’s homebrew ForceAtlas2 (by Mathieu Jacomy ). The first version of this algorithm was intended for graphs not larger than 10 000 nodes, while the Improved version of the Force Atlas 2 was made to handle up-to 1 million node large networks through approximation.

On this slide, you see the main component of the collaboration network for e-artexte.

## 14. Metadata visualization – co-authorship network (2)

With over six hundred thousand edges, we have to filter down. Limiting the network to those authors that have at least two co-authorship relations and showing only edges that have a weight of at least 5 (that means that we show only co-authorships that include at least 5 publications between the same two people), we get a more visually comprehensible graph with 420 nodes and 19 thousand edges.

## 15. Metadata visualization – co-authorship network – Sigma export – Select Hank Bull

Exporting this graph using the Sigma export makes it possible to explore and navigate authorship relations. As the user clicks on an author name, they see a list of all of the people they collaborated with. I mapped betweenness centrality to node size, and on the screenshot, you see that Hank Bull is visible with a high centrality, as he is positioned between two clusters of contributors. Hank Bull is a multidisciplinary artist, curator, organizer and arts administrator - and the fact that he shows up with a high centrality in the graph supports the statement that he holds fundamental importance for the development of visual arts in Canada.

## 16. Metadata visualization – artist network (1)

The e-artexte dataset also includes an artist field, where artists referenced in the texts are listed. Connecting two artists together whenever they are mentioned in the same item/exhibition creates a co-artist graph, if you will. There are items with over 100 artists listed. In the bottom you see what that looks like in the catalogue, a block of links that lead to lists of items. I include this in the slide because navigating through these blocks of links is difficult with no indications of importance for each name, which is the problem that visualization attempts to help solve.

The co-artist graph is very large with over 6.2 million edges

On the **left** is how the graph looks with some filters for edge weight and degree applied. Filtering that down further to focus in on the main “giant” component, and increasing the edge weight filter to at least 30 gives the graph shown on the **right**. Mapping betweenness centrality to node size provides for a visual presentation of the most influential artists in the collection.

##

## 17. Metadata Visualizations – Artists – Sigma.js (2)

The collective “General Idea”, along with Michael Snow are positioned with a significantly higher betweenness centrality, suggesting that they are influential in the sense of spanning across communities. Exporting this (3 thousand node - 42 thousand edge) graph of artists using the Sigma template shows this clearly and gives the user the ability to explore by navigating around the more significant (those with a degree of at least 100) members of the communities identified .

## 18. Metadata visualization – artists (3)

On this slide, we apply the Frutherman Reingold layout on just this small remaining set of most central artist relations in the collection. I believe that metadata visualization inspires serendipitous discovery. Who is at the center of the graph, and who is further out? Are there artist that I am unaware of that are connected to an artist I know? Is there a whole sub-structure within the graph that I am not familiar with?

## 19. Photography items – Keywords and Artists (1)

After some discussions of these results with Artexte librarians, we agreed that it would be particularly useful to create visualizations that combine keywords and artists in the same graph. One way to do that is to “cross” keywords with artists. So for each item, connect all the artists listed with the keywords listed.

The number of edges for the entire collection would be too large for Gephi to handle, so I decided to apply it to a data extraction from the catalogue, approximately 1800 or so items that are indexed with the “photography” keyword (in English or in French).

The resulting graph samples are shown on this slide, the unfiltered graph on the right, and a portion of the filtered graph on the left.

## 20. Photography items – Keywords and Artists - One Community (2)

The community detection and layout ends up showing keywords and artists together, which is a useful way to navigate this collection. Selecting a single community reveals the most important topics/keywords and their related artists.

 On the **left** you see the graph after a filter of edge weight of at least 2 was applied, which brings the number of edges down to under 100 thousand. Unsurprisingly, in the very center of the graph, there is photography and the french “photographie” - remember that this is a bilingual catalogue.

On the **right**, there is a close up of the community relatively distant from the center, shown in blue. Right in the middle of that community, one of the labels that stands out with a high centrality is Loui Emile Durandelle. He was a french photographer who took beautiful pictures in Paris at the turn of the century, for example, of the Eiffel tower under construction. While Artexte’s collection policy is to focus on contemporary visual art from 1965 onwards, the graph demonstrates that Loui Emile, a photographer working in the 1910s, was influential to photography in Canada.

##

## 21. Photography items – Keywords and Artists - Two Communities (3)

This slide shows a close up of the intersection of two communities on that keyword-artist graph, showing Architectural Photography in the middle, with other keywords and artists such as Lisette Model , the american photographer, who is positioned next to Walker Evans.

Examining this graph is a journey of discovery. Why is Lisette Model positioned next to Walker Evans? If you Google Image search Lisette Model, you will see her photographs, and you will also see a link to Walker Evans: it looks like Google Images also positioned them together (there is a little screenshot in the bottom right from the Google Images search). Interestingly, this only happens if you have your language setting in Google Images set on “French”. I suppose this is evidence that Google Images also uses graphs, and different ones depending on language - perhaps the bi-lingual e-artexte catalogue contributed to that relation between Walker and Lisette?

## 22. Photography items – Keywords and Artists - - Sigma Export - AI (5)

Exporting the graph to the Sigma template results in the ability navigate the graph, to select or search for a keyword, such as “Artificial Intelligence” and navigate from the Information pane to some of the related artists (like Karl Sims), and keywords (ike biology, evolution, interactivity, etc. )

## 23. Metadata as Open Data

This is the OR conference, and I should include some lessons learned and recommendations that I have for this community based on this work.

“The best thing to do with your data will be thought of by someone else”, I really like that quote.

Think of your metadata as the data, and the someone else as a digital humanities scholar trying to load it into a visualization platform.

I recommend that repositories consider whether they have the following barriers to access to their metadata:

1. Is XML or JSON export of **search results** technically supported on the public interface? If this is only supported with API calls, are these well documented to the public?

OAI-PMH, the only, shall we say, “consistent” method of extracting metadata from repositories, is great for some purposes, but it is not a particularly flexible method of access with its predefined sets. The option to export search results in XML directly from the public interface is available in EPrints, and Invenio – but not many other repositories.

In addition, a pre-generated, auto-refreshed (daily or weekly) downloadable metadata dump file for the whole collection would be useful. I realize that seems like a lot, but then again, we are supposed to be models of open access.

1. More specifically for metadata visualization: is an export format for Graphs supported?
	1. Support for multi-modal (bi-partite) graph format export using GML, an EdgeList file, or GraphML should be relatively simple to implement as it requires no projection
	2. RDF is a great and promising technology for semantic web, but it is not an easy format to work with for visualization.

Please consider exporting to a format that would be natively readable by a platform like Gephi, or Cytoscape. It would mean that a digital humanities scholar could simply export data from your repository right into one of these tools, without having to reformat using scripting languages.

**24. Metadata as Open Data - Access and Reuse Policy**

Lastly, but fundamentally important, is the metadata in your repository available for re-use? Is there a policy statement that allows download and re-use? The e-artexte catalogue has such a statement, granting permission for not-for-profit purposes.

## 25. University of Zurich - ZORA

Another recommendation is to experiment with incorporating network visualizations right into the repository yourself. In the EPrints community, a noteable example of this is the TrendTerms visualization by Martin Brändle, for the University of Zurich repository. This plugin extracts indexing terms from abstracts that are contained in the Xapian indexing engine used in EPrints, and constructs graphs of “trend terms” and their relations saved as graph XML files. These Graph files are then visualized with ProcessingJS.

In addition to this, ZORA repository includes an author collaboration graph for each deposit, showing a circular layout for collaborators of each item in the repository.

## 26. Useful Network Measures

To quote a really excellent overview article by Powell et al. called “Graphs in Libraries: a Primer”, “graphs need not be visible to be useful!” Calculating measures like degree or betweenness centrality can help improve the sorting options for retrieval based on “relevance”. The measurement of edge weight in a co-authorship network can be used to provide navigation to important related authors from an individual record display. Graphs can also be used as tools for identifying collaboration opportunities, to disambiguate author identities and aggregate related materials, to explore geospatial and temporal aspects of collections such as the changes in research focus over time.

You are all probably familiar with the COAR report on Next Generation repositories (from 2016).. Among the use cases listed in that report is the need for **Automated recommender systems for repositories,** and these require access to anonymised user-interaction logs, in the form of a triple. CERN repository has implemented this based on their own usage data, with recommendations of “similar articles”.

The release of access data would add another very useful dimension to the metadata graph.

## 27. THANK YOU

Going back to the e-artexte metadata visualization. The current status of this is that it has been made available to researchers in a local installation at Artexte, so that custom “on-demand” visualizations can be generated for researchers.

Next steps: Work with Artexte librarians to conduct usability studies around these visualization results.