

[Slide 1] How a Project to Digitize Legacy Paper Records Can Be Used as the Basis for a Transit System Knowledge Graph

TRB Annual Meeting, January 10, 2024, Washington, DC

Joseph Busch, Taxonomy Strategies

[Slide 2] Introduction

From 2020-2022, coincident with the height of the pandemic and the succeeding months of the “it's-best-to-work-from-home” waves of COVID surges, we engaged with a major regional transit authority, to help them reduce paper documents through digital scanning. A wide-ranging multi-jurisdiction railway, bus, and paratransit system, the Authority had undertaken this initiative to eliminate legacy paper records, to downsize their footprint while relocating 3,000 office-based employees that were spread across 10 aging buildings originally planned and configured nearly 50 years ago in the 1970's and 80's. The disparate facilities resulted in a fragmented workforce, and also had deferred maintenance and safety issues. The Authority was relocating most of these employees to 3 new office buildings.

[Slide 3] The goal of consolidating these employees to 3 regional buildings was to save money, and also hopefully to break down the silos and encourage collaboration among departments and teams. The challenge is that the new facilities are smaller office spaces, and the spaces are configured for new ways of working with office hoteling, hot desking, and privacy and collaboration areas. There will be limited space for personal materials such as paper files, engineering plans, and reference materials. To help plan for and prepare employees for the move to the new work environment, the Authority hired change

management consultants. Those consultants hired us to develop and execute the plan to more or less eliminate legacy paper records and reference materials. This involved developing the metadata strategy (aka taxonomy), doing the scanning, and building the repository that would become the system of record for legacy documents.

[Slide 4] Accidental Digital Transformation

But as the pandemic caused the Authority to make an abrupt shift to remote work, the focus of this engagement transformed into helping a bureaucracy which had no culture of remote work be able to work remotely.

[Slide 5] Enterprise Taxonomy Process in this Project

Although for enterprise taxonomies we normally use a top-down approach, in this project we were asked to adopt a **bottom-up approach** that would be executed iteratively department-by-department. In this process we:

- Surveyed each Department or business unit;
- Interviewed stakeholders in a departmental focus group;
- Defined use cases, metadata, and vocabularies (the taxonomy) for each department;
- Validated the taxonomy with the department against their use cases; and
- Generated a metadata specification for each department.

We applied best practices and guided departments toward emerging commonalities, but at the end of the day, each department got their own separate taxonomy.

[Slide 6] Document Survey

The type of documents that were held in the Departments varied widely. Paper documents in file folders or binders, including drawings of any size, were in scope for the project.

[Slide 7] Stakeholder Interviews

The agenda was primarily to discuss each type of document that was identified in the survey. The goal of the focus group was to identify for each type of document:

- 1) The primary way or ways that the department organized the documents,
- 2) The primary way or ways that the department searched for the documents, and
- 3) Typical use cases related to the documents.

Focus groups were conducted in person so that employees could show us examples of documents to help us understand their form and use. At the end of the focus group, we conducted a quick survey of the number of linear feet of materials, to estimate the number of pages of documents for planning the scanning effort.

[Slide 8] Draft Taxonomy

Following the focus group, we prepared detailed notes that were emailed to the participants for their review and correction. We used the Focus Group notes together with the Document Survey to prepare the first version of the Taxonomy. The Taxonomy identified:

- 1) The metadata attributes used by all Document Types in the department,
- 2) The Document Types and metadata attributes for each Type,
- 3) Any controlled vocabularies associated with metadata attributes,
- 4) The proposed file naming convention based on concatenating metadata values into three file naming segments,
- 5) At least one use case for each of the Document Types with the corresponding metadata that would fulfill each use case, and
- 6) The taxonomy to be used in the validation activity.

Following the focus group, we also prepared a presentation that summarized our preliminary assessment of the Document Types and our proposed metadata scheme for describing them so that they could be retrieved in response to the identified use cases. This “Discovery Report” was reviewed with the Department stakeholders in an online meeting. After the meeting, the stakeholders reviewed the report and provided corrections or approved it.

[Slide 9] Validation Activity

Based on the approved Department metadata scheme, an Optimal Workshop tree navigation exercise¹ was configured for each use case. Once launched, the department stakeholders had a week to complete the validation activity via a web link.

¹ This is called [Treejack](#) in Optimal Workshop.

Then we prepared a validation report based on the Optimal Workshop self-scoring activity results, and made any changes to the taxonomy that were warranted. The department would receive the Validation Report along with our final draft of the Taxonomy.

[Slide 10] The Specification

Finally, we prepared a full specification for how each departmental Document Type should be described with complete and consistent metadata. A QR-code control sheet was generated for each document type that would trigger the appropriate metadata template, so that technicians could generate the relevant metadata during the document scanning process. The specification also was designed to serve as a continuing specification, so that future scanning or born-digital documents could be appropriately tagged with the relevant metadata.

[Slide 11] Project Team

The project ran for 2 years from October 2020-October 2022, throughout the Covid-19 pandemic. I personally ran most of the focus groups, masking up and taking transit, meeting in conference rooms at various facilities across the Authority's area of coverage including in buildings in the train yards at the end of the rail lines. The rest of the strategy team was from [History Associates](#). [National Office Systems](#) (NOS) ran the scanning and tagging operation, and the change management vendor was [The Clearing](#).

[Slide 12] Project Demographics

We completed metadata strategies for 63 departments or business units from across the Authority. This included business functions ranging from accounting and HR, to bus and rail operations, to the transit police and health and safety departments. We completed an average of 2.6 department metadata strategies per month. There was an average of 9.6 attributes per department. The only mandatory attributes were Document Type and Business Unit for every department. Aside from the mandatory attributes, the most common attributes were:

Attribute Type	Attribute Name	%
Date	Date (MM/DD/YYYY)	59%
Name	Last Name	52%
Name	First Name	51%
Identifier	Contract#	25%
Identifier	Employee/Contractor ID	40%
Subject	Title	40%
Date	Year	37%
Location	Line/Station Code	25%
Name	Vendor/Manufacturer	25%
Date	Month/Year	21%

[Slide 13] Things, Not Strings

In 2012 Google introduced their knowledge graph using the tagline “things, not strings.” The key feature is that Google was now searching for named entities, that is, the names of people, organizations, locations, events, products, etc. and returning curated web content explicitly related to these entities rather than simply returning a list of web links as the search results. This required a pre-defined set of named entities, unique identifiers for them, and information that was explicitly related to them.

[Slide 14] Named Entity and Identifier Relationships

This is a reorganized list of the most common attributes in the transit system document collection from the digitization project. We've grouped them by named entities and identifiers, and indicated the linkages between them, that is, the identifiers that are associated with named entities. Since these named entities and identifiers have been associated with each document when it was scanned in the digitization project, actual documents are already associated with these entities. So, for any given entity or combination of entities, we can provide search results of explicitly related documents (related to things), not just those which happen to occur as a string in the text.

[Slides 15] Graph of Relationships between Employees and Policies

This is a graph of the transit agency Employee and Policy entities, and the relationships between them. Each named entity (or thing, which is called a "concept" in this tool) has attributes associated with it. These attributes are things like identifiers, titles, and dates. For example, an "Instruction" may have attributes such as "Instruction #," "Instruction Title," and "Effective Date."

There may also be relationships between entities. These may be hierarchical (called "specializations" in this tool) or associative (called "related" in this tool). For example, an "Instruction" is a specialization of a "Policy" which in turn is a specialization of a "Standard Operating Procedure" (SOP). An "SOP" may apply to a "Position" which is the job of an "Employee." An "Employee" is employed by a "Business Unit."

[Slides 16] Graph of Relationships between Contracts and Vendors

This is a graph of the transit agency contracts and vendors and the relationships between them. This graph resembles a business process model. The key difference is that the knowledge graph is about specifying named entities with attributes that facilitate the reliable retrieval of things. While the goal of the business process model is to specify a repeatable business process.

[Slides 17] Graph of Relationships among Transportation Assets

This is a graph of the transit agency assets and the relationships between them.

[Slide 18] RDF behind the transit agency asset graph

RDF is a serialized form of XML that is used to encode attributes and relationships for concepts (things) to communicate them between systems. To enable this every entity has a persistent identifier. These identifiers are maintained as one or more name spaces.

[Slide 19] Summary

In this talk we have provided a brief description of a project to digitize legacy paper records at a large transit agency. While the project generated separate metadata specifications for each business unit, an effort was made to guide the organization towards common practices where the same metadata attributes were used in multiple departments. In these cases, common naming conventions were applied as much as possible. At the end of the project we created a concordance of all the metadata attributes. This master list was used

to generate examples of graphs that could be globally applied at the transit agency for employees and policies, vendors and contracts, and transportation assets.

[Slide 20] Knowledge Graph vs. BPM

The goal of a business process model is to specify a repeatable business process. While a knowledge graph is about specifying named entities with attributes that facilitate the reliable retrieval of things.

[Slide 21] Questions

I'm Joseph Busch, principal consultant at Taxonomy Strategies. Do you have any questions?