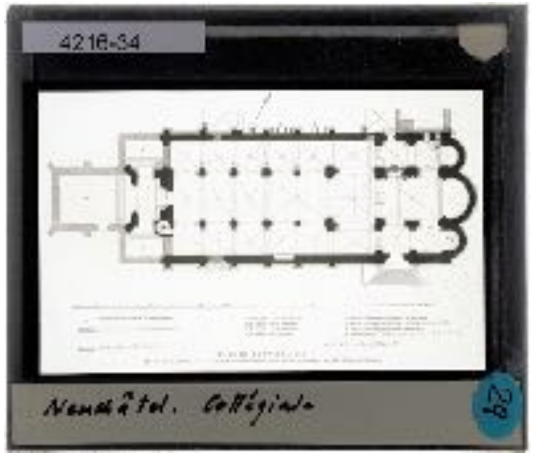
Overview Tuesday

- Questions
- CIDOC-CRM
- Extensions
- Afternoon: Setup and modelling of own data

Data Model

- Write down the classes and properties for a historic slide archive

Carouge S. [illegible] 10.
4207-25


4216-34
Neuchâtel. Collégiale
28

Eigentum des Kunstgeschichtlichen Seminars an der Universität Zürich

CIDOC-CRM

Introduction

A bit of history

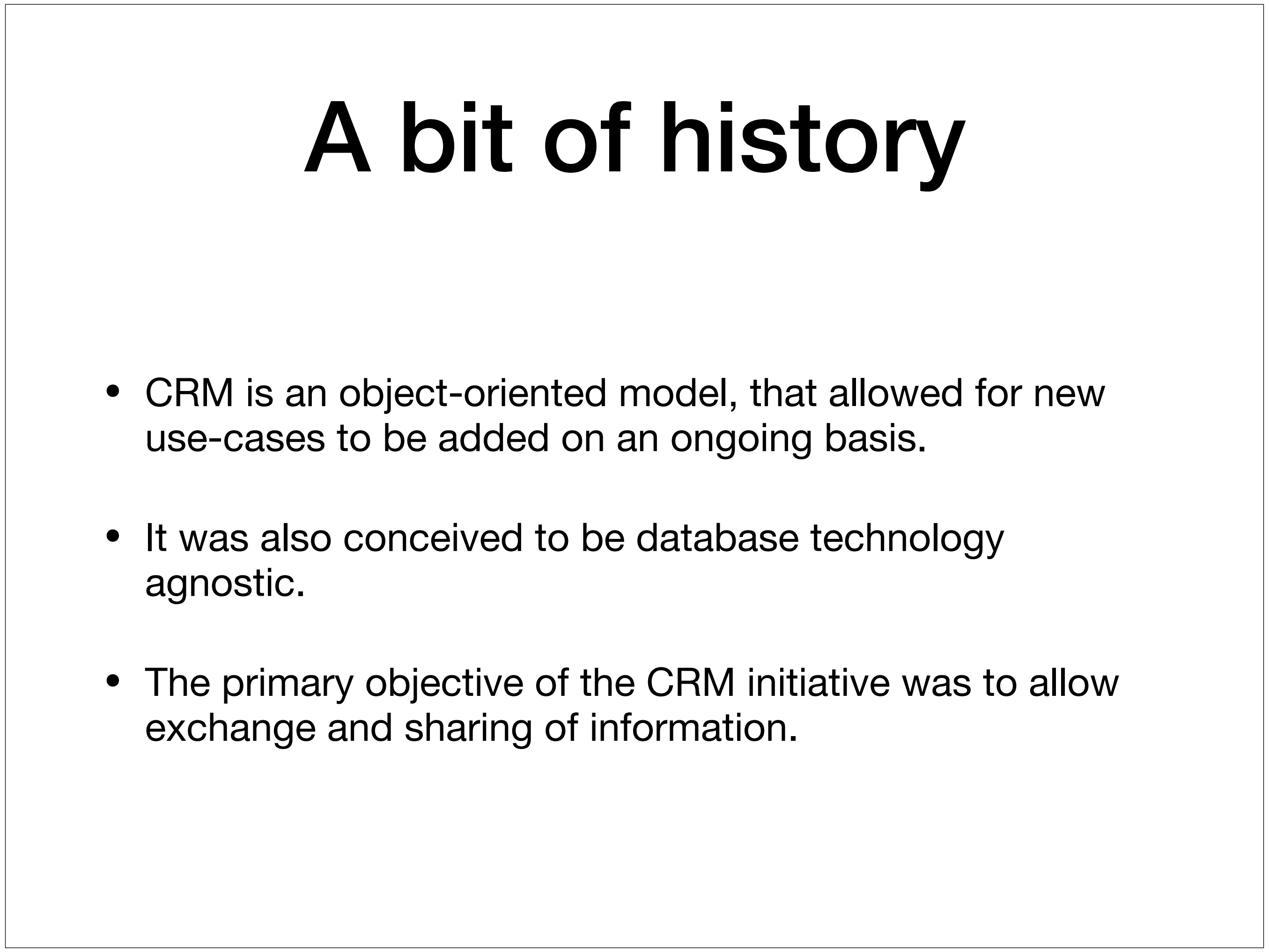
- Until 1998 CIDOC existed as an Entity Relationship model, being derived from the technology of relational databases.
- Not being flexible enough, it meant supporting a highly complex system that is impossible to maintain.
- In 1996 CIDOC-CRM was born as a project to replace the E-R-model.

Relational Databases are not good at relations

Name	ID
Peter	1
Susanne	2
Hans	3
Julia	4

Relation	ID
Mother	1
Father	2
Daugther	3
Son	4

Name_ID	Relation_ID	Name_ID
1	2	3
2	1	3
4	3	1
3	4	1



A bit of history

- CRM is an object-oriented model, that allowed for new use-cases to be added on an ongoing basis.
- It was also conceived to be database technology agnostic.
- The primary objective of the CRM initiative was to allow exchange and sharing of information.

CRM is an object-oriented model, that allowed for new use-cases to be added on an ongoing basis.

Object-oriented models are derived from programming practices in the 90s. What this means is, that an object has to belong to a class and the class itself describes the properties of a group of objects. A class can have sub-classes, which inherit all properties of the super- class and add additional ones. This makes sub-classes necessarily more specialized and super-classes more generalized. Objects can be polymorph, meaning that if they belong to a sub-class, they automatically also belong to the super-class.

It was also conceived to be database technology agnostic.
The primary objective of the CRM initiative was to allow exchange and sharing of information.

A bit of history

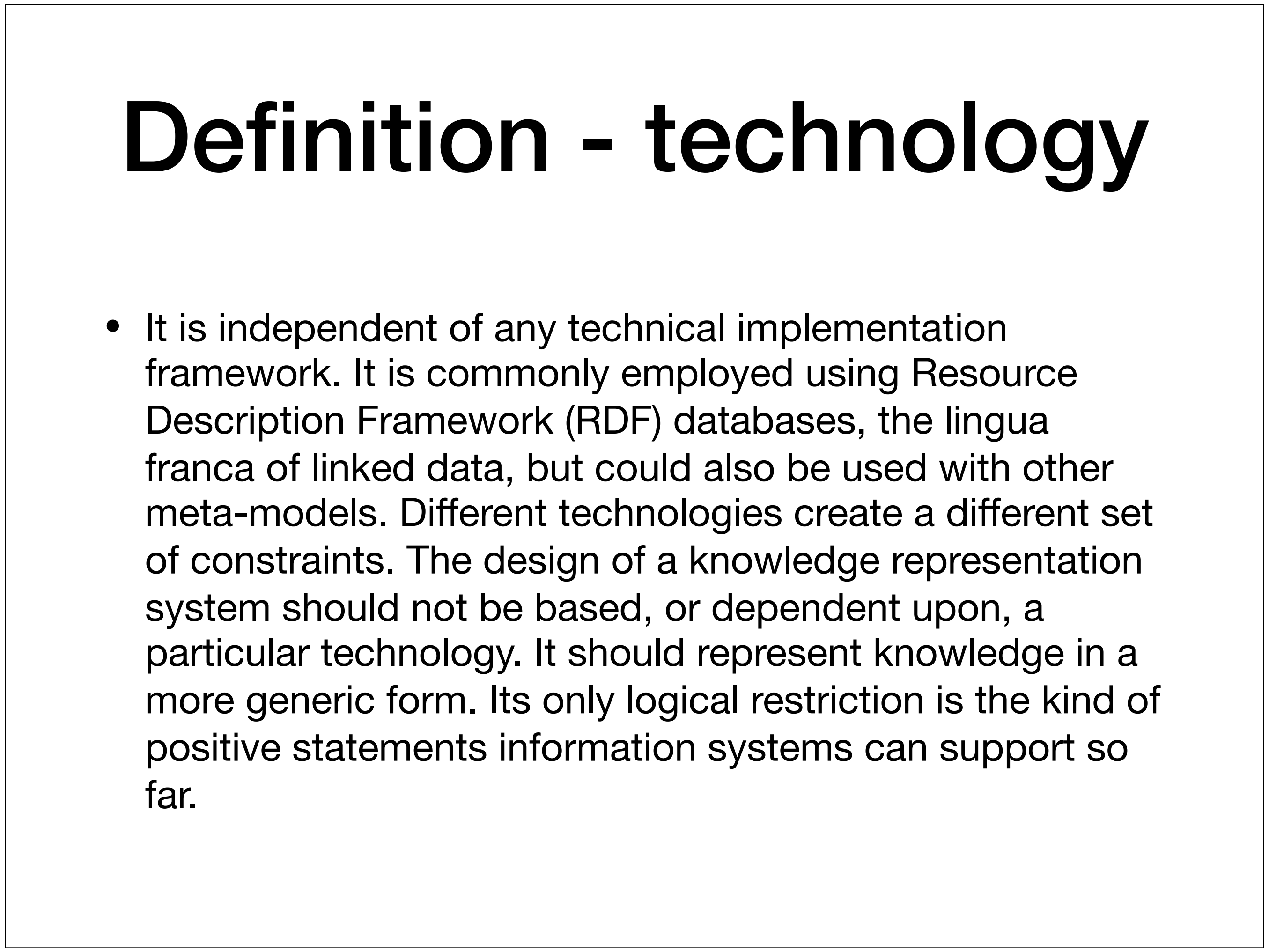
- Coming from a Computer Science background, automated reasoning was a big thing for the CRM creators. They define it as: „the ability to formally manipulate the data using logical rules in order to generate new information“.

- CRM clearly followed political topics: „In contributing to this resource of information institutions become important members of a revolutionary digital research community.“

- Or: „The value and relevance of data increases when it is communicated with its full meaning and context. This relevance is magnified when the knowledge of different institutions is combined to enable different perspectives to be preserved.“

Definition - general

- The CIDOC CRM is an ontology - a form of knowledge representation. An ontology represents the categorical knowledge within a domain, in this case the cultural heritage domain. The function of a domain ontology is to mediate the variability within a domain and provide a framework under which we can collaborate despite having different datasets. It is a language, not a statement of current scholarly convictions.

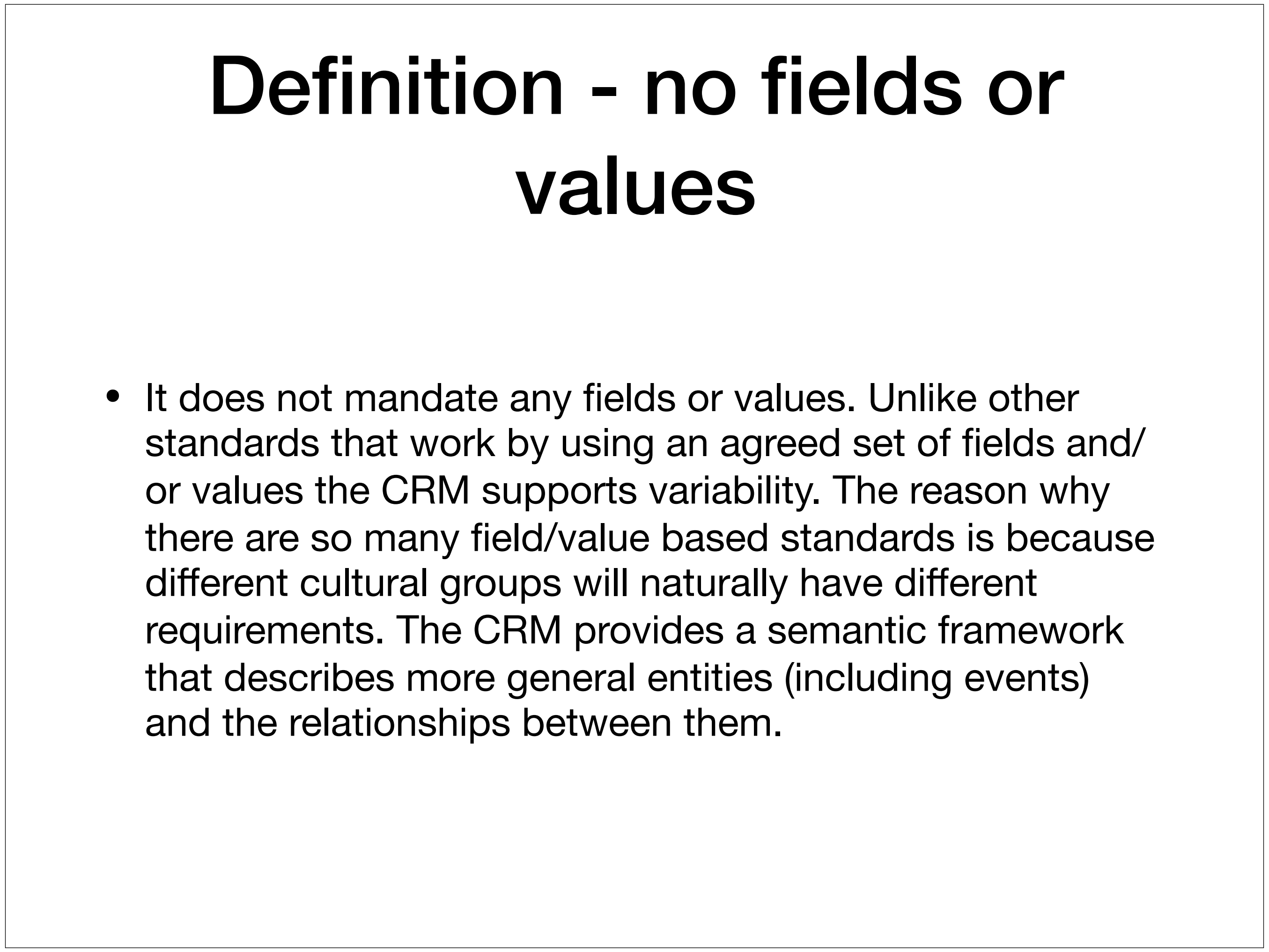


Definition - technology

- It is independent of any technical implementation framework. It is commonly employed using Resource Description Framework (RDF) databases, the lingua franca of linked data, but could also be used with other meta-models. Different technologies create a different set of constraints. The design of a knowledge representation system should not be based, or dependent upon, a particular technology. It should represent knowledge in a more generic form. Its only logical restriction is the kind of positive statements information systems can support so far.

RDF is a Standard model for data interchange on the web and a way for machines to work with facts.

RDF S P O
Linked Data S URI P URI O URI



Definition - no fields or values

- It does not mandate any fields or values. Unlike other standards that work by using an agreed set of fields and/or values the CRM supports variability. The reason why there are so many field/value based standards is because different cultural groups will naturally have different requirements. The CRM provides a semantic framework that describes more general entities (including events) and the relationships between them.

Rather than describing a limited number of common fields, as in many digital representations, the CRM describes objects more fully including the semantic meaning of the data. An aggregator may define a field called ‘description’ as a target for some object text, but organisations have different types of object descriptions that are created for different purposes, by different people and, therefore, may be interpreted and represented differently. These different perspectives are understood by organisations internally and may or may not be clear in digital collection management systems. Either way the CIDOC CRM provides the means to transfer that internal meaning and provide a more meaningful way of integrating data. General software developers are rarely museum documentation experts. It is far easier to define a set of fields based on a more superficial understanding and ask organisations to approximate to them. This creates overly generalised data integrations that have limited use. Essentially the representation of data has been left to groups that do not have the necessary understanding and without engagement from those who do. This is why true digital representation of cultural heritage data requires a positive collaboration with those with local and expert knowledge of the data. We wouldn’t produce a physical exhibition without providing the intellectual context for it, but we seem content to publish data without any concerns about its interpretation and context. This essentially undermines and erodes the purpose of cultural heritage organisations as centres of knowledge.

Definition - bottom up

- It is an empirically based ontology. Rather than being defined by a committee (top down), the CRM is based on empirical analysis of real practice and local knowledge (bottom up). The CRM develops as a result of understanding existing models of practice that have themselves developed over a considerable period of time; it represents nearly twenty years of international research. It is unlikely that a similar exercise would come up with a significantly different result. It is scientifically constituted and not influenced by the strength of opinion of a particular group or expert.

Definition - poly-hierarchical

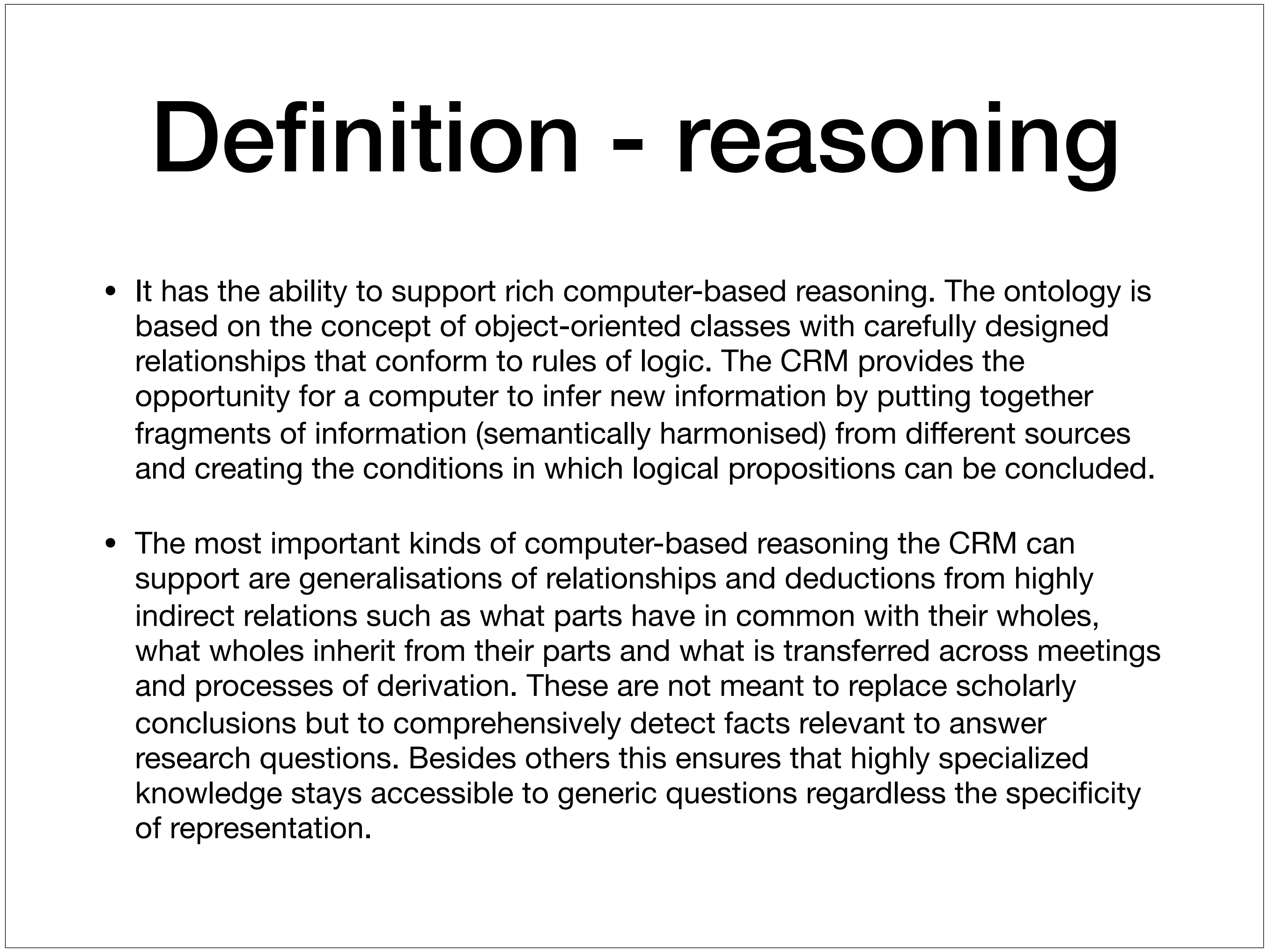
- It is poly-hierarchical (not a flat linear structure) providing an optimal range of generalisation/specialisation above the point of individual institutional terminological descriptions. In such a framework context and semantics become important.

Definition - terminology alignment

- It does not concern itself with differences in terminology between institutions, it supports the ability to “plugin” local terminologies and provides an ontological framework under which these vocabularies (conceptual terminology) can be compared and linked.

Definition - automation

- It provides a framework for matching instances of people, places, things, events and periods using the information and context around these entities. It does not need to rely on primitive string matching techniques.



Definition - reasoning

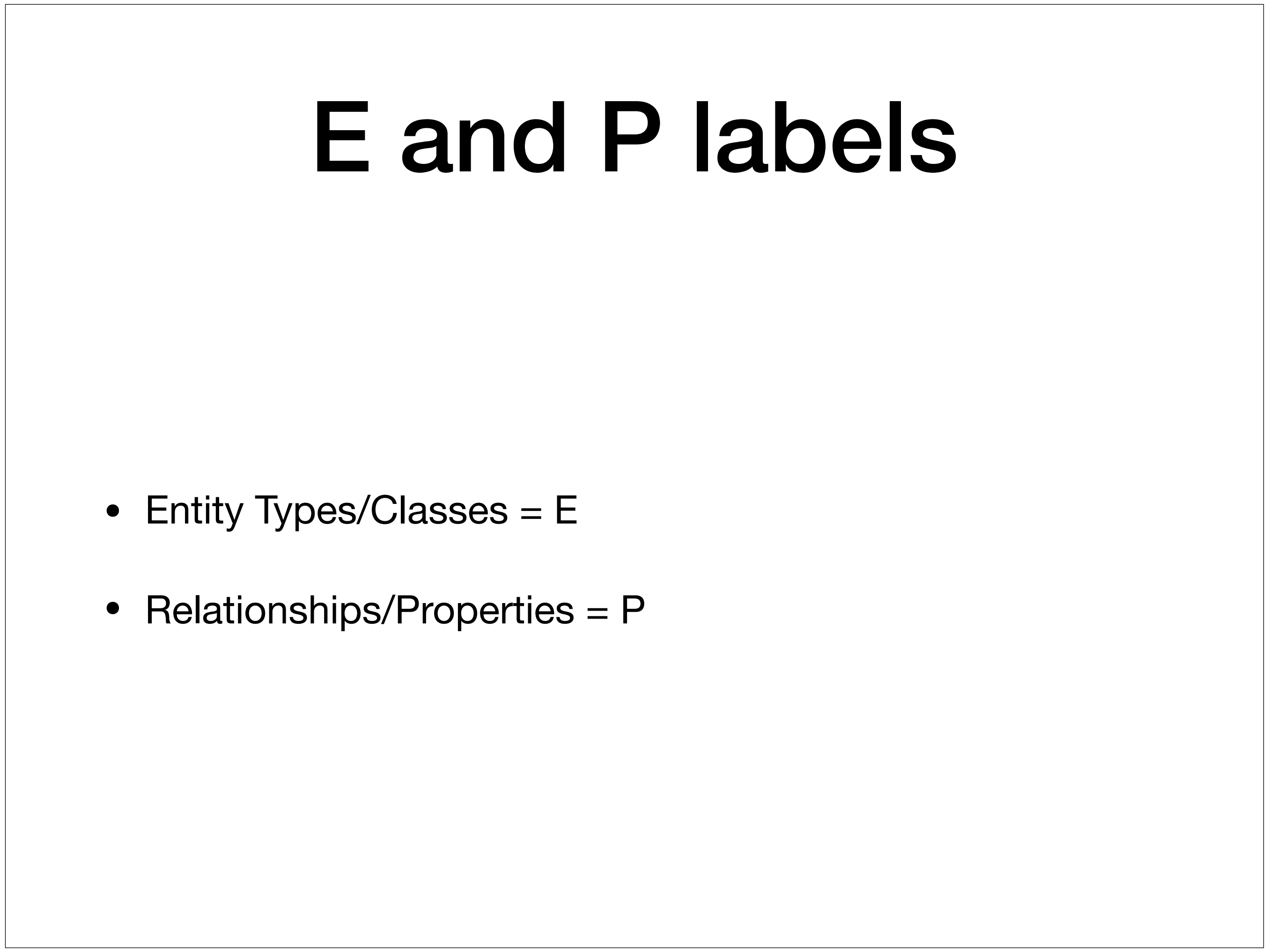
- It has the ability to support rich computer-based reasoning. The ontology is based on the concept of object-oriented classes with carefully designed relationships that conform to rules of logic. The CRM provides the opportunity for a computer to infer new information by putting together fragments of information (semantically harmonised) from different sources and creating the conditions in which logical propositions can be concluded.

- The most important kinds of computer-based reasoning the CRM can support are generalisations of relationships and deductions from highly indirect relations such as what parts have in common with their wholes, what wholes inherit from their parts and what is transferred across meetings and processes of derivation. These are not meant to replace scholarly conclusions but to comprehensively detect facts relevant to answer research questions. Besides others this ensures that highly specialized knowledge stays accessible to generic questions regardless the specificity of representation.

The relationships in CRM have been designed to support computerised reasoning but this ability is dependent on using relationships correctly and with the correct entity types. Therefore understanding the initial mapping process is very important. For example, the CRM relationship, “carried out by” can only be used between an “Activity” entity type and an “Actor” (Person or Group) entity type.

Mapping

- CRM acknowledges that somebody has to do the mapping of the data. Because if the source data has only implicit semantics, then a human (read domain expert) is needed to express this information explicitly.



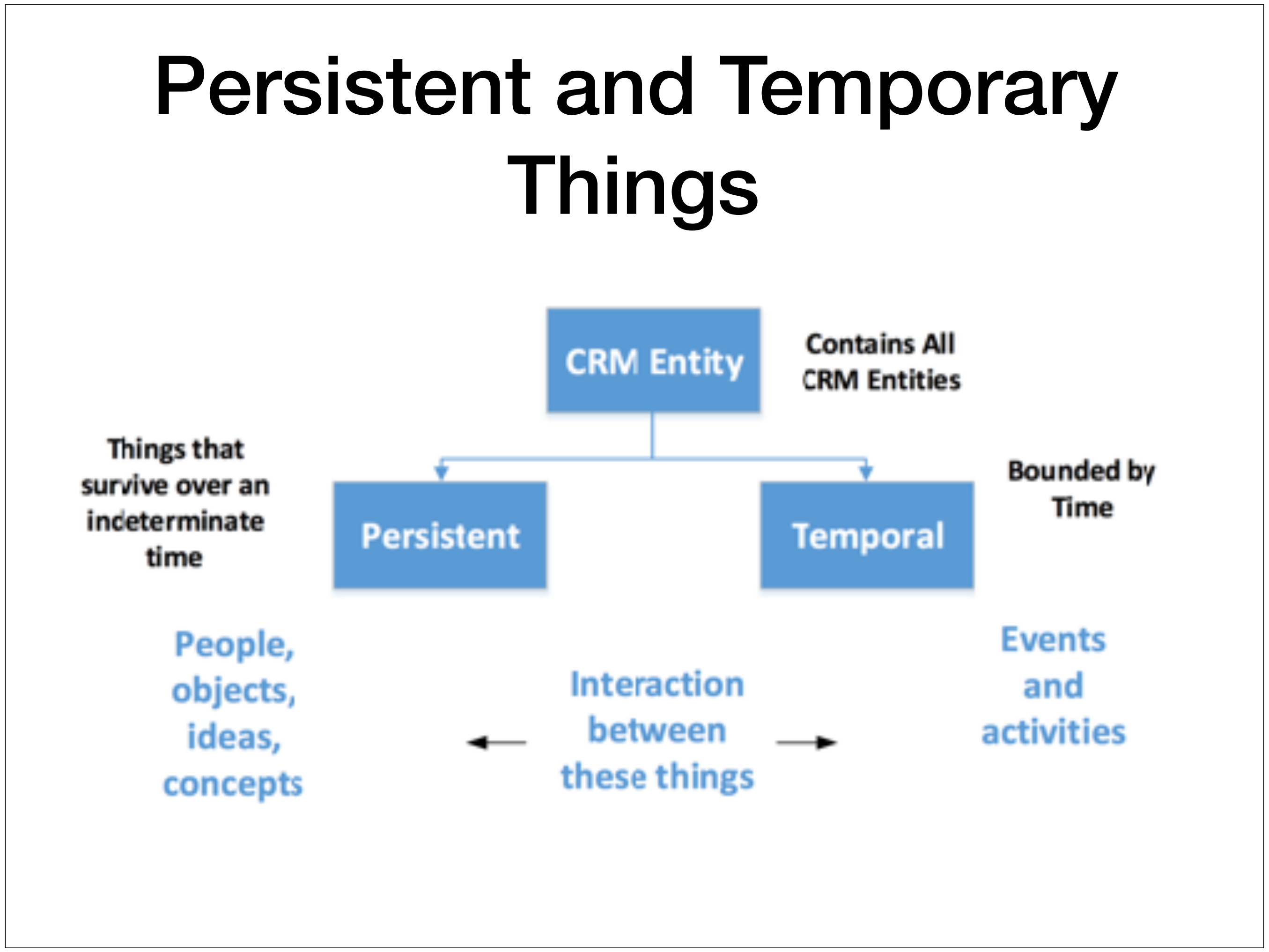
E and P labels

- Entity Types/Classes = E
- Relationships/Properties = P

The CIDOC CRM defines entity types and gives these short labels using the prefix of ‘E’. For example,
E22_Man-made_Object
Entity labels capitalise the first letter of each word.

Relationships (or Properties) use the prefix ‘P’, again with a label, for example, P1_is_identified_by
Property labels are lowercase.

Labels are convenient placeholders but the label does not (and cannot) convey the full meaning. They may be translated to other languages. Therefore the identity of the concept lies in the short label. To understand what any entity or relationship (property) actually means you need to refer to the scope note contained within the CIDOC CRM reference manual and understand the context in which they can be used. Along with the scope notes, the ‘domain and range’ (rules for what relationships can be used with certain entities) of the relationships, provide the information you need to map your data. Domain and Range are always just a convention, as depending on the direction of the property they might be inverted.



Persistent and Temporary Things

CRM Entity

Contains All CRM Entities

Persistent

Things that survive over an indeterminate time

Temporal

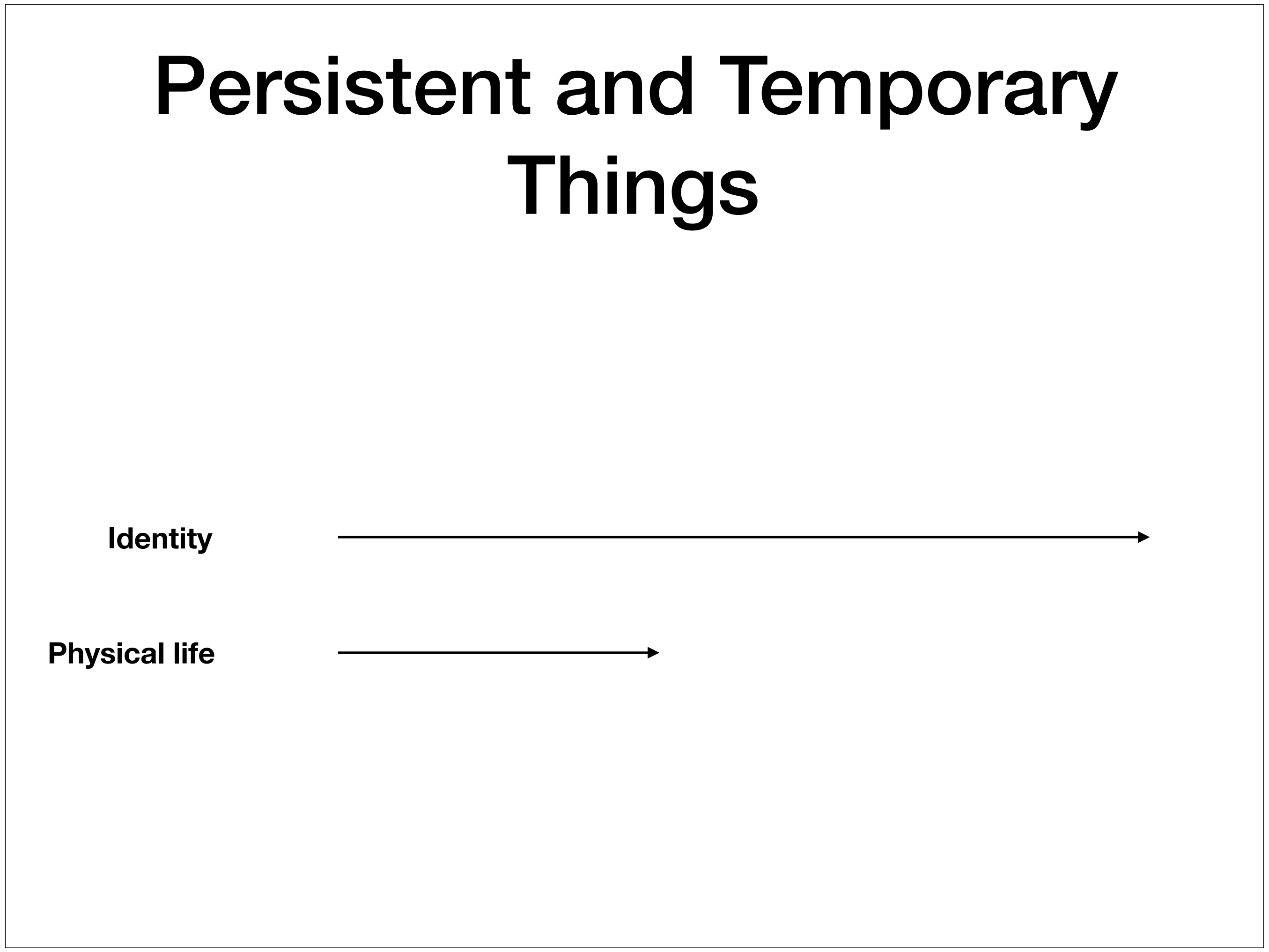
Bounded by Time

People, objects, ideas, concepts

Interaction between these things

Events and activities

There are the two key (disjoint) branches of the CIDOC CRM. There are things that have a persistent identity that can by their nature survive one or more events (physical things or ideas and concepts), and there are the temporal concepts (or phenomena) that have a nature of happening rather than being, over a limited time frame (an event or activity, like creation, or a historical period). Persistent entity types define instances that are initiators, recipients or witnesses of events and activities. They may originate, survive or terminate in events. This is the essence of the CIDOC CRM's event based model.

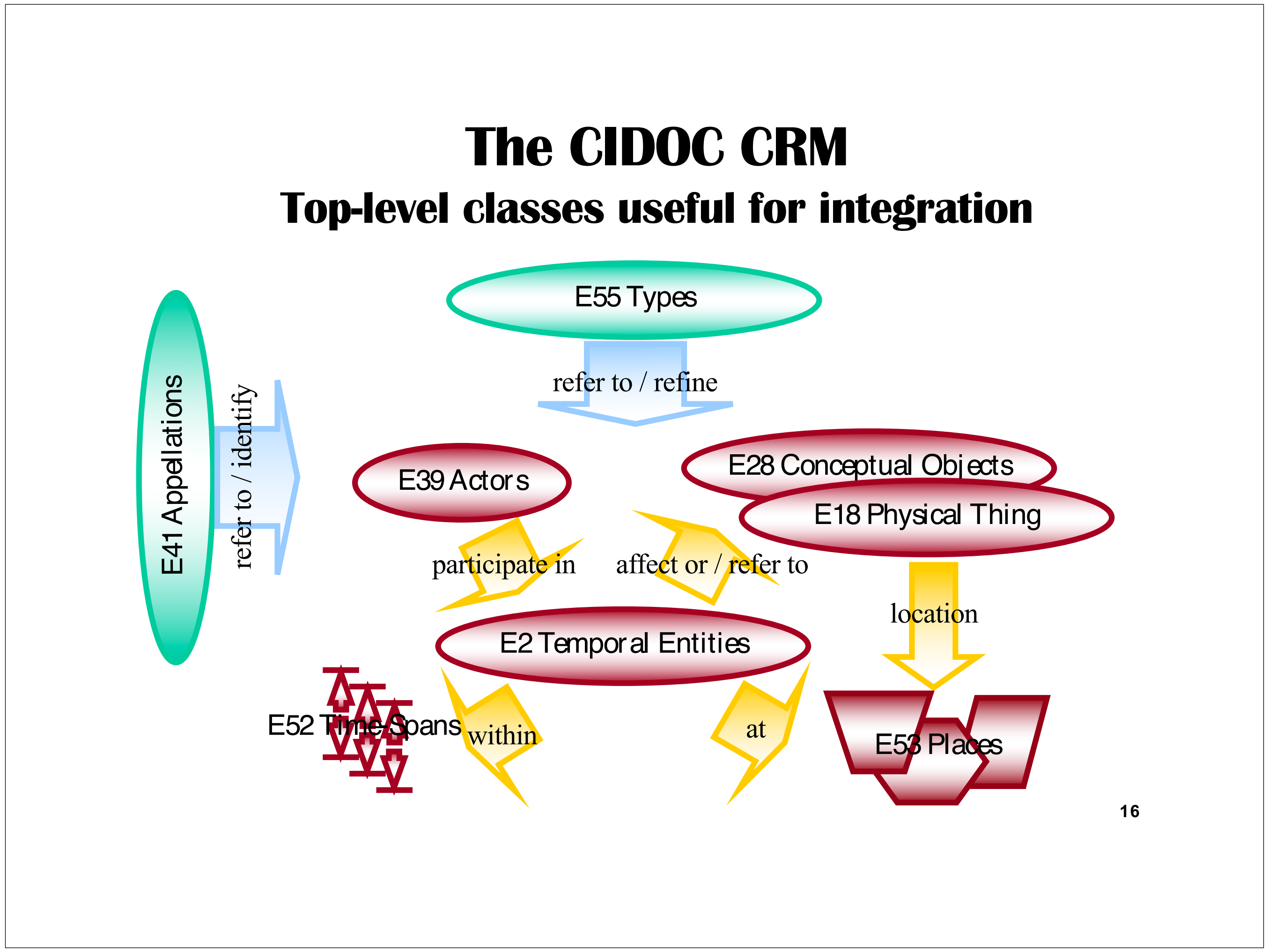


Persistent and Temporary Things

Identity

Physical life

For example, a person’s identity endures regardless of his/her death. Death is a temporal concept just as events that occurred during a person’s physical life are also examples of things bounded by some period of time. Leonardo da Vinci no longer exists physically but his identity survives. The sinking of the Titanic was a temporal event; although the ship sank to the bottom of the Atlantic Ocean, the identity of the Titanic lives on.



The CIDOC CRM

Top-level classes useful for integration

We talked about Top Level yesterday. If you understand this then you understand the CRM because everything else is a specialisation of this top level.

The CIDOC CRM is event based. At the core of this event model are Temporal Entities (E2) - things that have happened in the past.

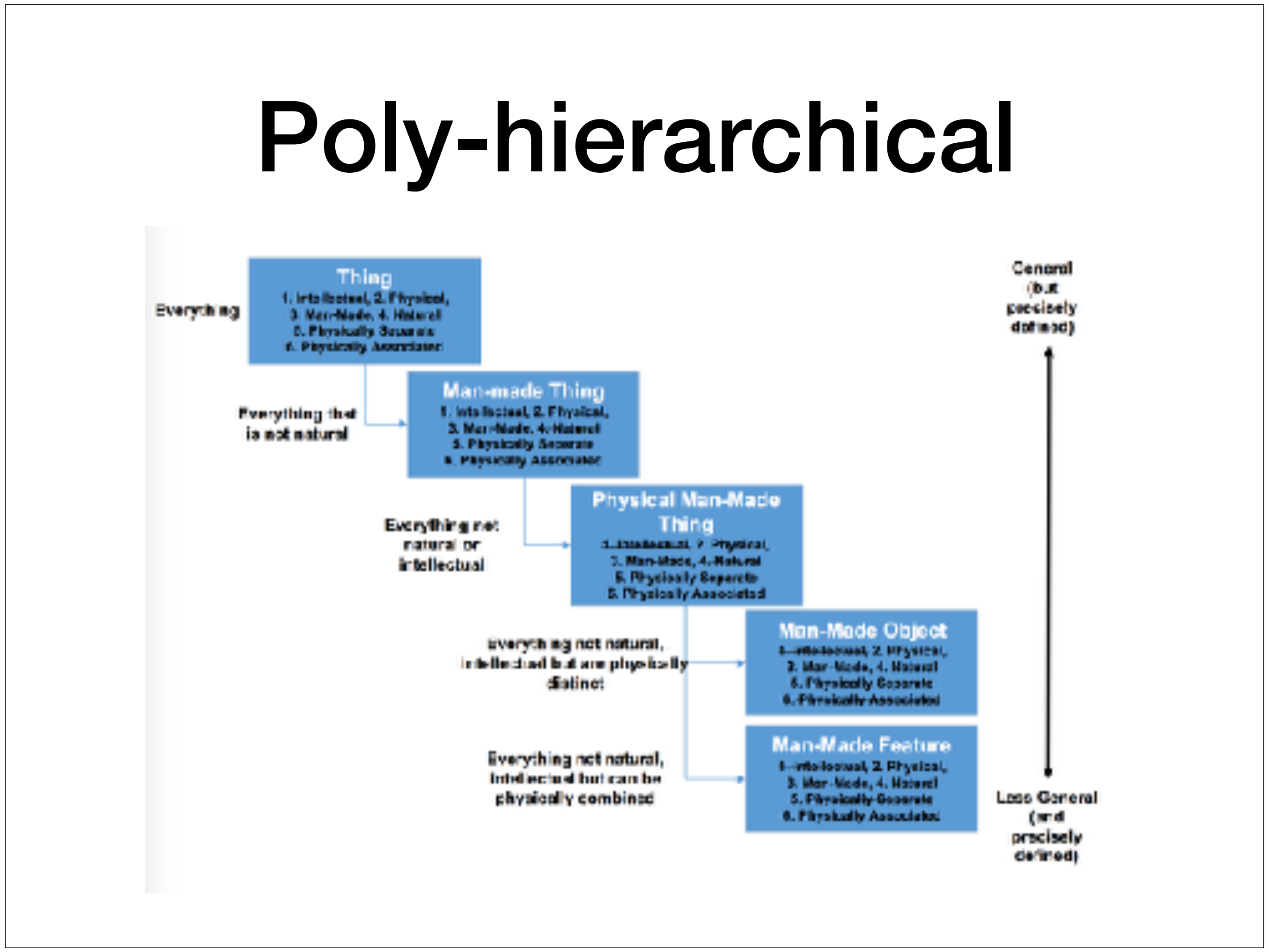
⍰ Only Temporal Entities can be linked to time and have Time Spans (E52). Objects (Conceptual (E28) and Physical Things (E18), Actors/People (E39), and Places (E53) cannot be directly linked to time. Therefore they must be linked to an event – a Temporal Entity (E2).

⍰ A Place (E53) could be a geographical location on earth, but equally it could be a location defined as the front of a ship or the inside of a ring. These are places that are geometrically defined.

⍰ Actors (E39) are entities with legal responsibility and an actor could be an individual or a group, for example, a school of artists or a company, and so on. Actors interact with things – both Physical Things (E18) and Conceptual Things (E28).

⍰ Physical Things (E18) are destroyed when they cease to be functional in the sense of our domain of documentation and therefore destruction is not necessarily linked to physically disappearing. A thing could be physically destroyed and transformed (created) into something else preserving parts of it. That new thing then becomes part of our domain of interest.

⍰ Conceptual Objects (E28) cannot be destroyed unless all carriers of it are destroyed. A carrier could be a book, a computer disk, a painting, etc., but it could also be the human mind. So destroying a conceptual object requires destroying all of its carriers, including people.



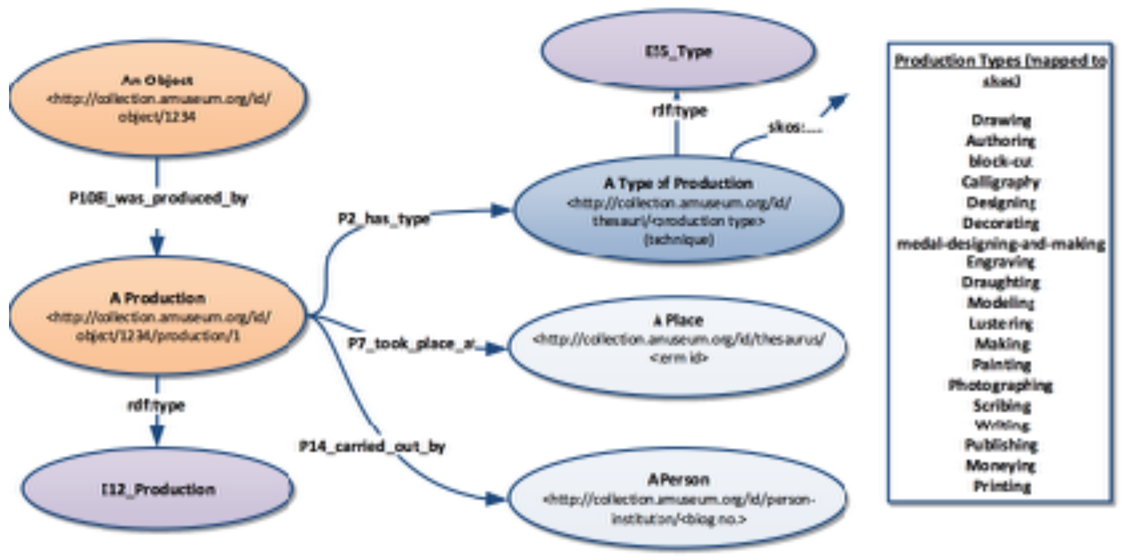
Poly-hierarchical

Entity types and relationships both exist in a hierarchy of meanings that provide different levels of generalisation (or specialisation depending on the way you look at it). This is important because we cannot always be precise about everything we want to describe, but when we can, we should. However, there is a point at which specialisations cease to become useful for harmonising data and where institutions might disagree. The diagram below shows how this hierarchy works. Entity types have sub types that are increasingly more specialised. Take a general entity like Thing. A ‘CRM’ Thing refers to things that have a stability of form, it could be man-made or natural, physical or intellectual, a feature of something else or a distinct object. If we know little about a Thing then we might use just this broader definition.

The key to understanding the CRM is understanding its structure - an understanding of the entity types and properties (and their descriptions through “scope notes”), the framework of relationships and how those relationships can be applied to the entity types, and applying these to an organisation’s understanding of their own data. This is not a technological undertaking.

By using events, CRM allows for making differentiated levels statements about individual entries.

Roles



For the harmonisation of data, this is the level of modelling that can be agreed upon. More specialized information, in this case the Type of the production event, is handled through a category Type and is usually pointing to a list.

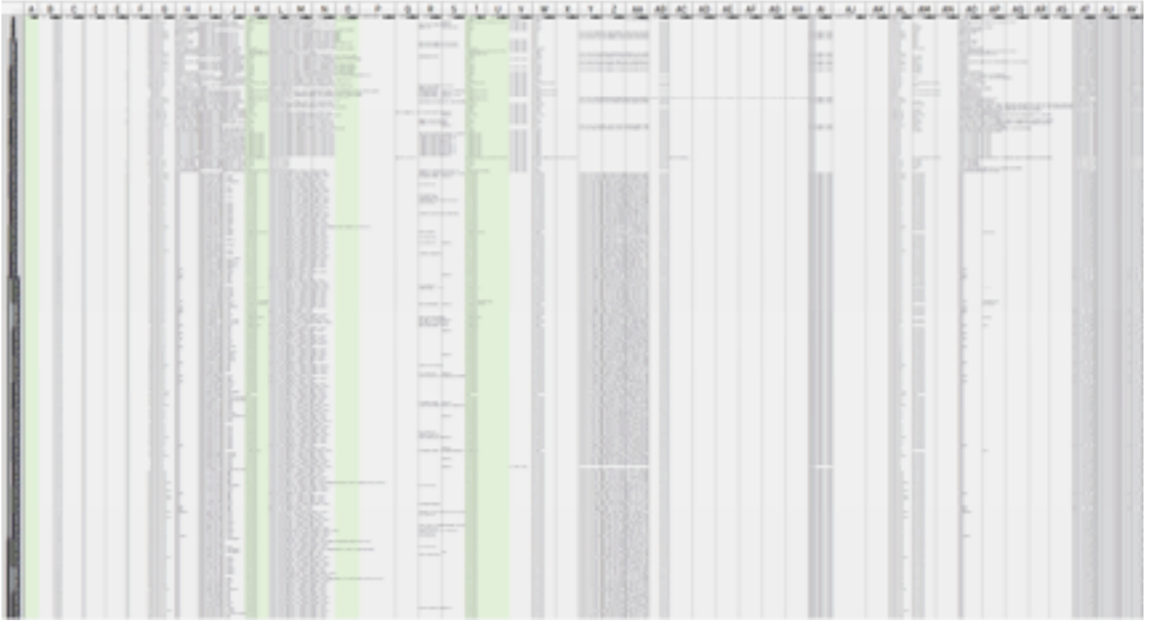
What data are we working with

- Manual, meaning human, data collection
- Data collection sometimes spread over several decades
- Data has had previous migrations and additional imports from external sources
- People entering data know more than the data shows

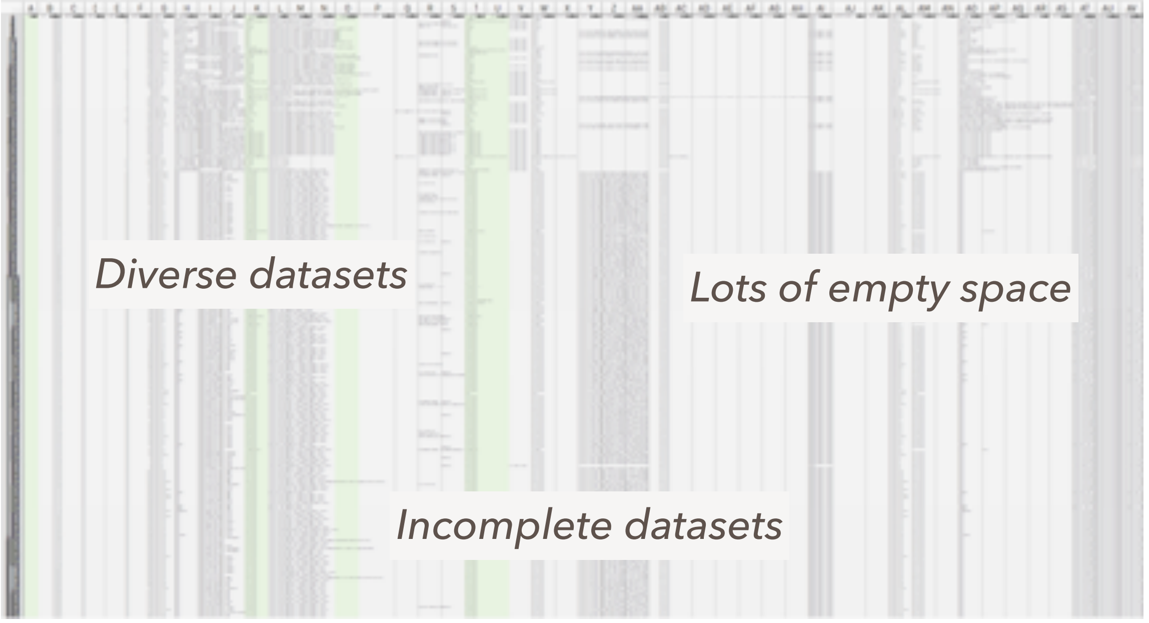
What data are we working with

- Manual, meaning human, data collection
  - **Different backgrounds, mistakes, bad days**
- Data collection sometimes spread over several decades
  - **Different computer literacy, changing leadership**
- Data has had previous migrations and additional imports from external sources
  - **Already messy data is becoming even messier**
- People entering data know more than the data shows
  - **Limitations due to data formats, user interface, lack of support**

Here is how it looks like



Here is how it looks like

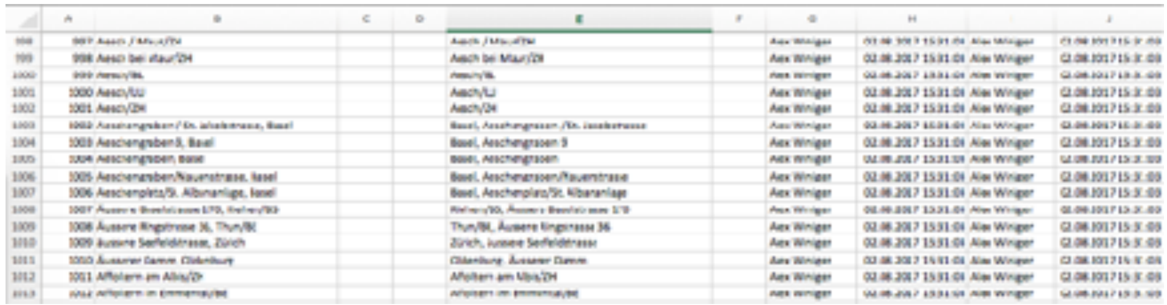


Diverse datasets

Lots of empty space

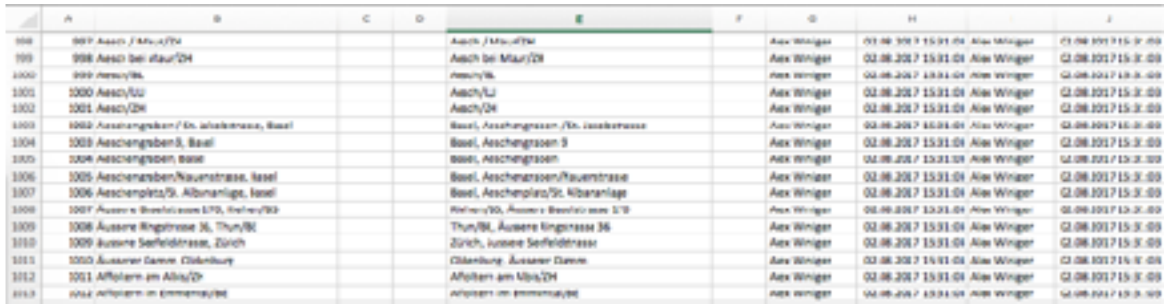
Incomplete datasets

Here is how it looks like



	A	B	C	D	E	F	G	H	I	J
998	997	Aesch / Maur/ZH			Aesch / Maur/ZH		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
999	998	Aesch bei Maur/ZH			Aesch bei Maur/ZH		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1000	999	Aesch/BL			Aesch/BL		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1001	1000	Aesch/LU			Aesch/LU		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1002	1001	Aesch/ZH			Aesch/ZH		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1003	1002	Aeschengraben / St. Jakobstrasse, Basel			Basel, Aeschengraben / St. Jakobstrasse		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1004	1003	Aeschengraben 9, Basel			Basel, Aeschengraben 9		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1005	1004	Aeschengraben, Basel			Basel, Aeschengraben		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1006	1005	Aeschengraben/Nauenstrasse, Basel			Basel, Aeschengraben/Nauenstrasse		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1007	1006	Aeschenplatz/St. Albananlage, Basel			Basel, Aeschenplatz/St. Albananlage		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1008	1007	Äussere Baselstrasse 170, Riehen/BS			Riehen/BS, Äussere Baselstrasse 170		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1009	1008	Äussere Ringstrasse 36, Thun/BE			Thun/BE, Äussere Ringstrasse 36		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1010	1009	Äussere Seefeldstrasse, Zürich			Zürich, Äussere Seefeldstrasse		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1011	1010	Äusserer Damm, Oldenburg			Oldenburg, Äusserer Damm		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1012	1011	Affoltern am Albis/ZH			Affoltern am Albis/ZH		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1013	1012	Affoltern im Emmental/BE			Affoltern im Emmental/BE		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03

Here is how it looks like



	A	B	C	D	E	F	G	H	I	J
998	997	Aesch / Maur/ZH			Aesch / Maur/ZH		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
999	998	Aesch bei Maur/ZH			Aesch bei Maur/ZH		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1000	999	Aesch/BL			Aesch/BL		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1001	1000	Aesch/LU			Aesch/LU		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1002	1001	Aesch/ZH			Aesch/ZH		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1003	1002	Aeschengraben / St. Jakobstrasse, Basel			Basel, Aeschengraben /St. Jakobstrasse		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1004	1003	Aeschengraben 9, Basel			Basel, Aeschengraben 9		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1005	1004	Aeschengraben, Basel			Basel, Aeschengraben		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1006	1005	Aeschengraben/Nauenstrasse, Basel			Basel, Aeschengraben/Nauenstrasse		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1007	1006	Aeschenplatz/St. Albananlage, Basel			Basel, Aeschenplatz/St. Albananlage		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1008	1007	Äussere Baselstrasse 170, Riehen/BS			Riehen/BS, Äussere Baselstrasse 170		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1009	1008	Äussere Ringstrasse 36, Thun/BE			Thun/BE, Äussere Ringstrasse 36		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1010	1009	Äussere Seefeldstrasse, Zürich			Zürich, Äussere Seefeldstrasse		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1011	1010	Äusserer Damm, Oldenburg			Oldenburg, Äusserer Damm		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1012	1011	Affoltern am Albis/ZH			Affoltern am Albis/ZH		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03
1013	1012	Affoltern im Emmental/BE			Affoltern im Emmental/BE		Alex Winiger	02.08.2017 15:31:04	Alex Winiger	02.08.2017 15:31:03

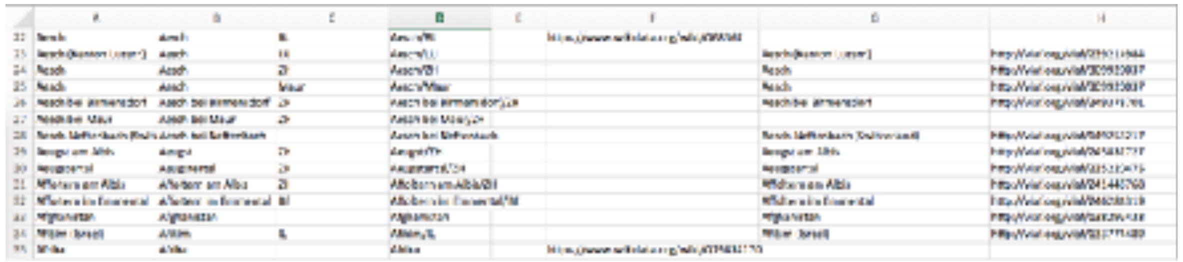
Typos

Vague statements

Conflicting entity types

Overloaded data fields

Data Reconciliation



	A	B	C	D	E	F	G	H
22	Aesch	Aesch	BL	Aesch/BL		https://www.wikidata.org/wiki/Q[illegible]		
23	Aesch (Kanton Luzern)	Aesch	LU	Aesch/LU			Aesch (Kanton Luzern)	http://viaf.org/viaf/[illegible]
24	Aesch	Aesch	ZH	Aesch/ZH			Aesch	http://viaf.org/viaf/[illegible]
25	Aesch	Aesch	Maur	Aesch/Maur			Aesch	http://viaf.org/viaf/[illegible]
26	Aesch bei Birmensdorf	Aesch bei Birmensdorf	ZH	Aesch bei Birmensdorf/ZH			Aesch bei Birmensdorf	http://viaf.org/viaf/[illegible]
27	Aesch bei Maur	Aesch bei Maur	ZH	Aesch bei Maur/ZH				
28	Aesch-Neftenbach [illegible]	Aesch bei Neftenbach		Aesch bei Neftenbach			Aesch-Neftenbach [illegible]	http://viaf.org/viaf/[illegible]
29	Aeugst am Albis	Aeugst	ZH	Aeugst/ZH			Aeugst am Albis	http://viaf.org/viaf/[illegible]
30	Aeugstertal	Aeugstertal	ZH	Aeugstertal/ZH			Aeugstertal	http://viaf.org/viaf/[illegible]
31	Affoltern am Albis	Affoltern am Albis	ZH	Affoltern am Albis/ZH			Affoltern am Albis	http://viaf.org/viaf/[illegible]
32	Affoltern im Emmental	Affoltern im Emmental	BE	Affoltern im Emmental/BE			Affoltern im Emmental	http://viaf.org/viaf/[illegible]
33	Afghanistan	Afghanistan		Afghanistan			Afghanistan	http://viaf.org/viaf/[illegible]
34	[illegible] (Israel)	[illegible]	IL	[illegible]/IL			[illegible] (Israel)	http://viaf.org/viaf/[illegible]
35	[illegible]	[illegible]		[illegible]		https://www.wikidata.org/wiki/Q[illegible]		

Extracting other entities

Identifying correct entities

Special characters require UTF-8

Returning unclear places

Contrary to statistical analysis, we are interested in the whole set.

Your data model

- What kind of classes did you write down? And what kind of properties would you define?

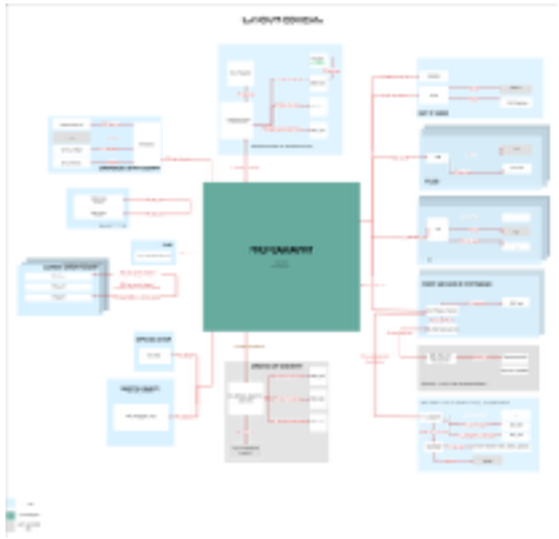
- How did these definitions came about?

Result of First iteration

Clear focus on the workflow

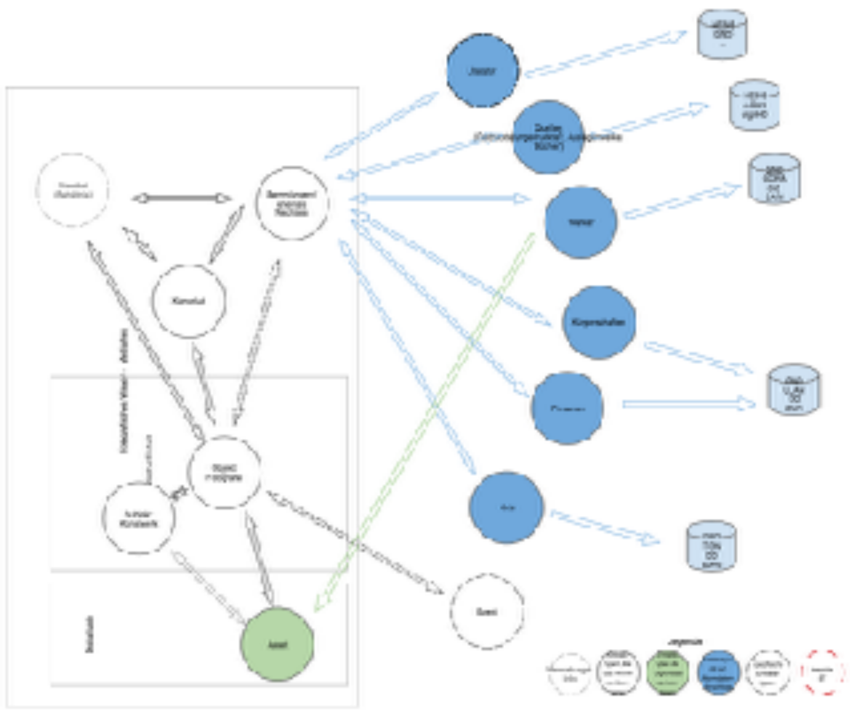
Current iteration

Reusing standards

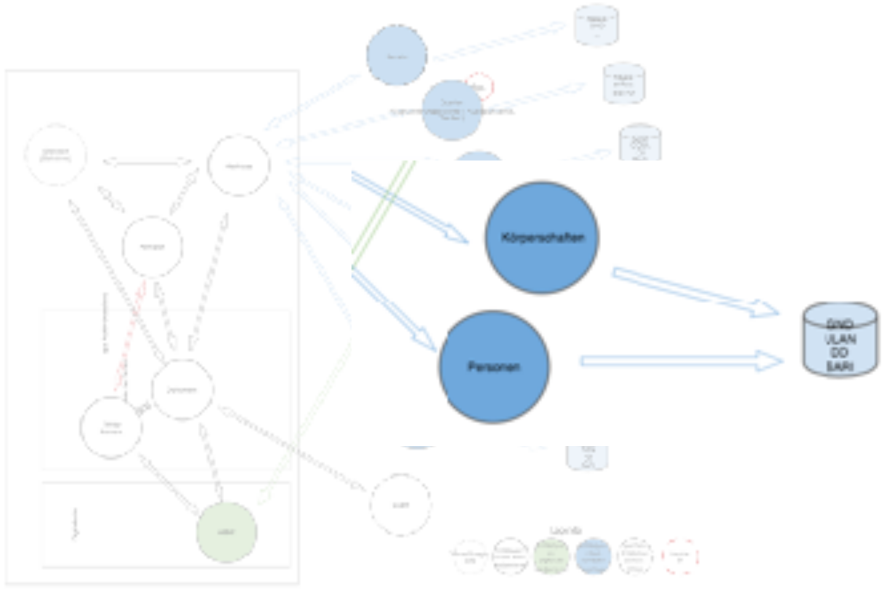


…

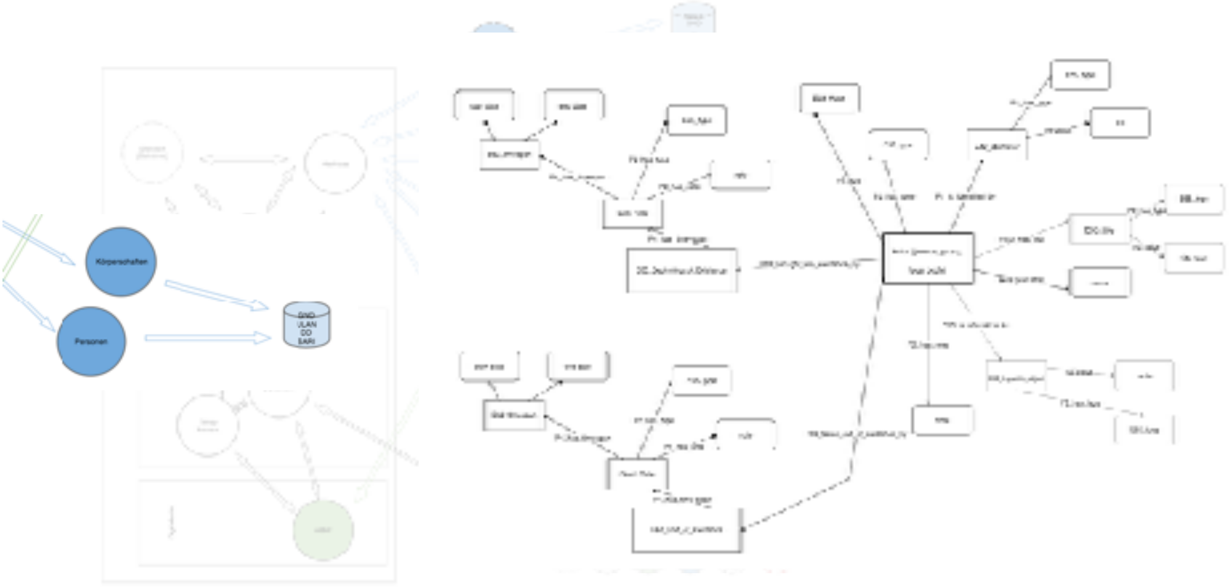
Example



Example



Example


Körperschaften
Personen
GND
ULAN
DD
BARI