The ‘Linking Persons’ project, an outline

1. **Introduction**
   In order to be relevant, the result of the ‘Linking Persons’ project should be made extensible. As soon as a firm basis of acceptance has been established concerning the validity and feasibility of the approach advocated in the present document, we should be able to demonstrate its applicability to datasets beyond our span of control. We believe we will achieve such extensibility by following as closely as possible the design principles underlying DBpedia’s Databus (https://www.dbpedia.org/resources/databus/). Any project follow-up based on this Databus approach may expect to have its path cleared in a significant way if we could tap the Databus know-how accumulated in the bosom of DBpedia’s international community. Therefore, we shall from the very start write our texts in English and share them on this international platform.

2. **Our modus operandi**
   This project is all about linked open data: historical data should be made easily accessible to the general public wider audience and, moreover, should be structured the way life is structured: even the most remote of connections might be relevant and, therefore, one should be able to get there by just following a path. Out of its belief in the value of high-quality data that is publicly accessible and guaranteed not to serve some commercial goal, it is common for the open data movement on the internet – to which the initiators of this project belong - to contribute qualified work for free. For this reason, the ‘Linking Persons’ project is a project that needs no funding and no budget.

   All the same, its added value should be measured according to well-defined criteria. Of these criteria, two factors stand out:

   * The drive of its initiators, and the time they want to make available to the project;
   * The cooperation of a handful of organisations in the field of cultural heritage. Such cooperation entails, that these organisations are willing to support and sustain this project with a little time spent by members of their staff. Also, it would be helpful if such cultural heritage organisations would put to the disposal of the initiators any digital datasets deemed to be relevant to the aims set by the project.

   Unless the project would ever scale up, no financial support is required. Probably not even for the phase in which we would connect to the DBpedia Databus in order to align our approach to other cultural datasets ‘out there’. Equally, however, there will be no deadlines to the project. It will be considered finished as soon as the aims of the project can be judged to be proven or disproven, both by its initiatiors and by the cultural institutions involved.

3. **What is this project about**
   The ‘Linking Persons’ project is a volunteer project by two members of the Dutch Chapter of the DBpedia Association, Gerald Wildenbeest and Gerard Kuys. Our aim is to provide a set of linked open data on some historical subject, thereby offering a handful of practical proposals on how linked data in the field of cultural heritage ought to be structured.

   Such proposals are meant to offer, in the field of cultural heritage, an alternative to datasets

that have their models very much derived from the collectioner's view: the past as a process that has delivered items to be described, and a process that has in an inimitable way made sure some items ended up in some cultural heritage collection while others did not. The alternative that we intend to deliver, is a structure for cultural heritage datasets that goes beyond the boundaries of collections essentially established from the early 19-th century onwards [1]. Such a structure should do two things: facilitate annotations from people not in the cultural heritage world, and offer a general underlying structure for corroborating assertions about subjects so diverse as the genuineness of source material or unexpected connections of one phenomenon with another.

This is not altogether what we see in the endeavours undertaken in the cultural heritage world, unfortunately. The use of linked data (in its most universal technical form, RDF) is limited to implementing a kind of mapping of one data entity type onto the other. If it is to be no more than that, we could simply map the description of a particular collection of materials onto the other – and would hardly need the inherent qualities of RDF to do exactly that. In such a case, interoperability of one collection with the other would mean that, in order to tell a consistent data story about some historical person, we would be obliged to map collection scheme A onto collection schema B and possibly also onto collection scheme C. This really is interoperability, of course, but much more can be done. We do feel, that a linked open data representation of things past should not depart from the perspective of material items that for some reason have withstood the tooth of time. Being organised on the basis of collection items, the former approach is fragmented, and tends to reproduce the divisions between the cultural institutions (based on books, archives, recordings, artifacts etc.). Worse even, it tends to extend these divisions onto our understanding of things past.

Rather, we would prefer our data models to start from the assumption, that differences in perspective should remain with the beholder, and not be inherently present in the data themselves. And, since almost every practitioner of history advocates a different past, we should organise our data as a 'discussion without an end' (Pieter Geijl). Linked open data, concerning anything in the past that is deemed to be worth our while, must facilitate the juxtaposition of different perspectives, must offer the best possible features for documenting (and challenging) the sources on which such views need to be based, and must facilitate comparison.

Seeking not to engage in a discussion on the merits of a collections-oriented approach of open data versus a more holistic approach, we set out to make our point by proof of concept with genuine historical data. By way of trying to solve a series of problems that is closely connected to reconciling a set of linked data vocabularies and fit them logically together, we intend to demonstrate that a logical infrastructure for a coherent and possibly general approach to historical data can indeed be constructed. Would we be the first to enter this uncertain field? What kind of models are vailable so far?

4. Holistic models for cultural heritage data – a fallacy

There is no general way of modelling changing entities over time. The underlying models of ever so many datasets reflect such diversity quite strongly. Frequent is the approach to model, like in a finite state machine, previous states of previous forms of some entities. Even so, the meaning given to 'finite state' is all but consistent. The lists of previous states turn out to be defined very differently by just as many datasets. This would not really be a

1 Cf. Marita Matthijsen, Historiezucht. De obsessie met het verleden in de negentiende eeuw, Vantilt Nijmegen, 2013.

problem, if we could map the various approaches to any common denominator, expressing the very core that matters when we try to describe how an entity – let alone a state of affairs – has evolved over a shorter or longer period of time.

To make this kind of mapping succesful, requires that we have a minimum of common understanding what the past (and Time) exactly is, and how to model it. What is certain, is that the past cannot be modelled in a way where the entities remain the same, having just their properties undergo an update. Almost all entity types are subject to the impact of time, and have a life cycle of their own. And not just these entity types in isolation, but also the relations they have with their surrounding world, and the time-dependent typologies they refer to.

In this project, we intend to inspect a little more closely the models that offer a handful of concepts about which linked data model renders best the digital representation of entities changing over time. Like any representation, such a model will never be complete, but we think we can now move towards a comprehensive way of modeling. Such endeavour can only be succesful if images of history – and therefore their digital representation – will be modeled according to a single set of guiding principles. These principles should be based on three ideas: the idea of integrality (there are discrete fields of interest concerning the past, but common to all are a particular time window, perhaps also a common spatial frame, and a common point in the development of society), the idea of the layeredness of experience (a tidal wave may be 'lived' in very much the same way by very different people, whereas famines, wars, or technological developments may have a quite different impact on one group of people compared to another), and, finally, the idea of the past as an experience that is to be remodeled every day, since the 'meaning' of history has a tendency to be interpreted quite differently every time some important development has broken into our own lives.

Rather than embracing a comprehensive approach addressing the issues concerning change over time, we prefer to base ourselves on a limited set of rules of thumb. The Piercean approach to knowledge representation, for example, as published in the *Knowledge Representation Practionary* by Michael K. Bergman [2], for our purposes is too general and maybe also too ideosyncratic. For Pierce and Bergman an event is a 'junction of states', a view we decided not to share. Wholesale approaches do not do what we want to do: presenting a showcase based on a dataset with real historical data, and in this way learn what exactly we need to know.

5. Principles to be tested

Let us pragmatically sum up the features that our data model must or might be able to provide. These are, respectively:

* Shared identities for people who have (had) a more or less proven existence in the real world;
* Views of history as collections of observations by some author, observations that must be based on digital documentation accessible to all;
* Shared vocabularies for expressing the relations between individual humans (family, (in)formal groups, etc.);
* Shared frameworks for the denomination of place and time;

2 Michael K. Bergman, *A Knowledge Representation Practionary. Guidelines Based on Charles Sanders Peirce*, Cham (Springer) 2019, blz. 134-135.

* Shared typologies of professions and their development over time;
* Shared typologies for the things, works, institutions or ideas bequeathed to posterity;
* Shared ways of documenting the origin of assertions and attributions about things past (metadata);
* A layered model in which search and disclosure, assertions about historical Persons, Places Objects etc., documentation to underscore assertions, documentation about the source and context of some assertion, and contemporaneous and documentary references, are all separate sections (layers, if you please) within the same model;
* Shared models of events: their scope (personal or societal), scale/granularity, frequency;
* Calibrate the overall framework. Get as close as we can to the structure of the upper ontology DOLCE+DnS Ultralite (DUL). Having Persons in some Role participate in Events is a model that resembles the A2A model structure of *Wie Was Wie* quite well. DOLCE+DNS Ultralite should also be our vehicle for expressing historiographic interpretations. We expect that the DUL design pattern of Information Objects might fulfill reasonable well our need to model assertions about a state of affairs in the past.

Since our project is about human history, Persons are bound to be the central category of our model. We would like to combine separate domains of discourse within a single model, thus getting closer to a model that integrates both personal relations between persons (kinship and the like) and relations of persons to places and to societal events. Given the fact, that in this way we risk ending up again with an all-encompassing model, we must integrate our historical data e*x posteriori.* This means that we should extend our horizon no farther than a given dataset with historical data (and available other datasets with historical data) allow us to link data about historical persons. We should stop at the point where we might have provided a framework for a digital representation of a tiny little bit of history. So, of the features summed up above, let us just work out those that need to be done in order to present the case of one particular dataset.

6. Modelling showcase data, measuring the model

We wouldn't have dared to propose a linked data showcase in this way, if we weren't certain that the cultural heritage data do exist with which our integrationist endeavour might have any chance of success. As to contemporaneous data about persons, there is always the risk of infracting limitations concerning these persons' privacy. Therefore, we should go back in time, thus further limiting the number of eligible datasets.

Fortunately, there is a dataset that represents a 19th-century encyclopedical work connecting (a 19th-century view of) persons to places, and sometimes also gives a short account of events that took place in the location under discussion. About fifteen years ago, the 13-volume work of A.J. van der Aa's *Aardrijkskundig Woordenboek* [3] (2nd edition, 1839-1851) has been OCR'd and made digitally searchable by the Centraal Bureau voor Genealogie (CBG) [4]. Additionally, an index of all person names in Van der Aa's text has been compilated by Mrs. Vennik-van der Linden on the request of the Rijksdienst voor het Cultureel Erfgoed. Tempting as it was to convert this material into linked data, no project was defined to do so, until one of us did a small-scale Proof of Concept for the CBG five years ago. This Proof of Concept limited itself to the town of Goes, and was looking how to match Goes-related persons in the civil registry *Wie Was Wie* (based on the XML A2A

3 http://nl.dbpedia.org/resource/Aardrijkskundig_Woordenboek_der_Nederlanden

4 Google did the same thing, but, due to poor transcription of 19th-century typography, the resulting file is hard to process (http://books.google.com/books/about/Aardrijkskundig_woordenboek_der_Nederlanden).

model). The Goes-match was extended with persons in *Van der Aa* on one hand and with person-related lemmas in DBpedia on the other.

We intend to carry on with this approach, and apply it to a larger dataset. But also, we aim to widen the scope both with regard to the way data about historical persons could be linked, as well as trying to place these persons in their proper settings both in place and in time. For the first 100 pages of Van der Aa's work, and for the relevant persons in Vennik's index, the linked data already exist – meaning also, that there is now a linked-data model for referrals in historiographical works. We propose to proceed as follows:

* Convert the descriptions of locations in Van der Aa to RDF, and publish them – probably in Triply on the NPLD platform during the phase of testing & commenting, and in a more final form as a contribution accessible by way of the DBpedia Databus framework. This Databus framework is going to provide us with a common infrastructure for provenance data;
* Link wherever possible the location description in Van der Aa to a representation of the locality in either DBpedia, *www.gemeentegeschiedenis.nl*, or, if possible, to historical content in the Basisregistratie Grootschalige Topografie (BGT)*;*
* Connect each locality to a type in a taxonomy of location types (especially, types of land use – from feudal rights to landed property, the government of which has grown to be increasingly profit-oriented since the times of the physiocrats. For the sort of mapping we want to do, see the table below);
* Identify persons wherever mentioned in any of Van der Aa's location descriptions (in Vennik's index, in DBpedia wherever possible, in *Wie Was Wie* if born later than, say, 1815). Find additional entries for one and the same person, even if the name is different. The link to *Wie Was Wie* wil allow to establish the 'true' name through certificates of birth.
* Connect each person to events related in the location descriptions, build up an event typology as far as Van der Aa's text is concerned;
* Connect persons in Vennik's index one to another from the perspective of kinship, as far as Van der Aa's work allows, using a standard genealogical vocabulary for linked open data [5];
* Wherever applicable: The form required to complete such a step, is to generate assertions with an author and a timestamp. In a 'discussion without an end', every contribution must be identifiable and traceable.

7. Why historical data on the Twente region

Obviously, completing all the steps mentioned in the previous paragraph would be far too much work when applied to the whole body of Van der Aa's 13-volume work and to all 23,000 persons in Vennik's index. Therefore, we restrict our scope to the Twente region, which can be considered to be a more or less closely-knit, easily identifiable community over a stretch of at least five centuries. And from a practical point of view: both authors have a personal link with the Twente region – the former lives there, the latter has written a dissertation on the mid-20th-century textiles industry in this region. And what is equally important, the regional museum De Museumfabriek (known before as Twentse Welle) in Enschede has a reputation for presenting the regional past from an integralist perspective on the history of Twente. Also, the Museumfabriek participates in the Netwerk Digitaal Erfgoed. Obviously, there is a common ground on which both parties – the Museumfabriek

5 Our preferred vocabulary for genealogical relations would be Robert Stevens's Family History Knowledge Base (http://ceur-ws.org/Vol-1207/paper_11.pdf)

and the linked open data zealots - may come to a fruitful collaboration.

8. **Describing a moving target**
As to the Twente region, we shall provide the following kind of mapping of current municipalities and localities to the geographical entities in Van der Aa:

Locality	DBpedia	Central geo-point	Description actual	URI-code in vd Aa RDF	Start page vd Aa text	End page vd Aa text	Locality type in vd Aa
Agelo	http://nl.dbpedia.org/resource/Agelo	52.3889, 6.8783	Buurtschap in de gemeente Dinkelland	351, 352	1-61	1-61	Groot Agelo – hamlet Klein Agelo hamlet
Albergen	http://nl.dbpedia.org/resource/Albergen	52.3717, 6.7619	Dorp in de gemeente Tubbergen	86, 457	1-73	1-73	Albergen - hamlet
Almelo	http://nl.dbpedia.org/resource/Almelo	52.35667, 6.6625	Stad en hoofdplaats van de gemeente Almelo	493, 494, 495, 496, 497, 498, 499, 500	1-98	1-100	Stad Almelo – town and municipality Ambt Almelo – municipality Almelo - Arrondissement Almelo - Canton Almelo - Ecclesiastical circle Almelo-en-Vriezenveen fief ('heerlijkheid')

Table 1: Mapping present-day localities to Van der Aa's 19th century locations (Twente region)

There is more to this than just mapping geographical named entities, however. Whereas the table above quite naturally departs from a point of view that covers the structure of modern administration and government, Van der Aa's 19th-century view was more differentiated. Apart from the municipal level – counting many more municipalities than we have now – van der Aa includes in his account the ecclesiastical organisation (which of course for each religion follows different principles), and the organisation of law enforcement in arrondissements and cantons. Also, there is wide variety in forms of land ownership. In Van der Aa's 19th century, vestiges of old seigneuries existed next to manors, estates and other semi-feudal ties to the land. How will we be able to map such subtleties onto today's straightforward maps?

9. **Cooperation with cultural institutions in Overijssel and in the Twente region**
We need clearly identifiable partners in the field of cultural heritage. For the permission to use a data section for Twente from the *Wie Was Wie* dataset, we will address a request to the Regional Historical Centre for Overijssel in Zwolle for permission to use a Twente-section of *Wie Was Wie* data. For De Museumfabriek, we hope to find common points of interest as well. We shall be needing their help, for example, for correctly identifying genealogical relationships, like the ones mentioned by Van der Aa and those (on the Stork and Scholten

families) described by a present-day authors like Jaap Scholten and Wim Nijhof [6].

10. **Modelling standards used**
    * SKOS for search-enabling structures only (taxonomies, thesauri, mapping of terms etc.);
    * DBpedia ontology;
    * Person Names Vocabulary (ING Huygens, Nationaal Archief). This vocabulary is exemplary in its biographical approach rather than modelling person data from the perspective of autorship or ownership;
    * The Places-in-Time vocabulary, which is the model underlying the ErfGeoViewer that resulted from the Erfgoed-en-Locatie project (2013-2014);
    * Geo-standards with layering, to be applied for several kinds of jurisdiction;
    * A genealogical standard, like the Stevens' Family History Knowledge Base;
    * Simple Event Model (Vrije Universiteit)|
    * DOLCE+DNS Ultralite (DUL) upper ontology

11. **Deliverables and future perspective|**
    At the end of our project, we will have a linked open data representation of the Twente-related items in A.J. van der Aa's *Aardrijkskundig Woordenboek* that allows us to query:

    * Van der Aa's decription of (the history of) 19th-century localities in what is now Twente, including descriptions of the structure of local ecclesiastic organisations and local law enforcement districts;
    * The mapping of these localities to present-day geographic information;
    * The list of persons who were mentioned in Van der Aa's account of the history of a locality in the Twente region (we expect 200-400 persons mentioned) ;
    * The mapping of well-known persons in that list to lemmas in DBpedia;
    * The matching of any persons mentioned with civil registry information in *Wie Was Wie*;
    * Identifying redundant person identities in Vennik's index;
    * Relating person information to events related in Van der Aa;
    * A taxonomy of events as far as Van der Aa's text goes;
    * For musea, the nice thing about linked open data is, that they may jump in real-time from one narrative to another. If 'Van der Aa's linked data' were to be added, there is a lot more to jump to, in this case to a body of information objects about the Twente region;
    * A piece of solid modeling of data on the human past, which - thanks to its upper ontology - is compliant with similar historical datasets published on the web.

6 Jaap Scholten, Horizon City, AFDH Enschede/Doetinchem 2014; Wim H. Nijhof, J.H. van Heek (1873-1957), kunst katoen en kastelen, Waanders Zwolle 2008.