SudocAD: A Knowledge-Based System for the Author Linkage Problem

Michel Chein, Michel Leclère, Yann Nicolas

SudocAD: A Knowledge-Based System for the Author Linkage Problem ⋆

Michel Chein* and Michel Leclère* and Yann Nicolas**

(*)LIRMM-GraphIK (CNRS,INRIA,UM2) (**) ABES

Abstract. SudocAD is a system concerning the author linkage problem in a bibliographic database context. Having a bibliographic database $\mathcal{E}$ and a (new) bibliographic notice d, r being an identifier of an author in $\mathcal{E}$ and r' being an identifier of an author in d: is that r and r' refer to the same author ? The system, which is a prototype, has been evaluated in a real situation. Compared to results given by expert librarians, the results of SudocAD are interesting enough to plan a transformation of the prototype into a production system. SudocAD is based on a method combining numerical and knowledge based techniques. This method is abstractly defined and even though SudocAD is devoted to the author linkage problem the method could be adapted for other kinds of linkage problems especially in the semantic web context.

1 Introduction

The issue addressed in this paper is a linkage problem that arises in bibliographic or document databases. The fundamental problem to be solved can be abstractly stated as follows. Let B and B' be two computer systems (e.g. bibliographic databases), let r be a referent in B to an exterior world entity and r' be a referent in B' to an exterior world entity (e.g. identifiers of authors of documents), then are r and r' referring to the same entity (e.g. the same author)? For a long time (cf. e.g. the seminal paper [NKAJ59]), various problems having linkage problems at their core have been extensively studied under different names (cf. [W.E06] and [W.E08]), such as:

- *record linkage* (e.g. [GBVR03],[Gom02], [EIV07])
- *entity resolution, reference reconciliation, co-reference resolution* (e.g., [BGMM$^+$09], [Gom02], [SPR07])
- *de-duplication* (e.g. [dCLGdS12])
- *merge/purge* (e.g. [HS98])
- *entity alignment*(e.g. [SE13], [MAS12])
- *object identification* (e.g. [SD05])

These problems have been considered for many kinds of databases, especially data warehouses where it is important to know if two identifiers refer to the same object or to different entities in the exterior world. These problems have increased in importance due to the web, especially with respect to Linked Open Data, where a key issue is to gather all available information concerning the same entity.

Let us mention some reasons that highlight the importance of linkage problems in bibliographic or document databases. Firstly, all of the previously mentioned problems are important in libraries: some of them are induced by the evolution of libraries (e.g. adding records to a base, merging bibliographic bases, maintenance of different bases), others concern the quality of the record bases (e.g. consistency inside and between bases, relevance of the subject). Secondly, a bibliographic database is a very rich source of information on the documents themselves and also on authorities. International work is under way to standardize the metadata (FRBR [frba], CIDOC CRM [cid], RDA [rda]), to build shared ontologies. This allows the transformation of a document base into a knowledge base and then knowledge-based techniques can be used to make the most of the information present in the base. Thirdly, whenever semantic web languages are used for solving object identification problems and due to the large size of document bases, the techniques developed are not only interesting *per se* but also as a testbed for linkage problems in the web of data context.

Most solutions to linkage problems are based on classification techniques (e.g. [FS69], [NRST09], [Gom02]). In such an approach, an entity is described by a list of attributes; attribute values are simple data (e.g. names, dates, numbers, etc.) and approximate similarity measures are assigned for each kind of attribute; a similarity measure is built for lists of attributes (often it is a weighted combination of the attribute similarities); finally, a decision procedure allows to state when two lists of attributes represent the same entity. Recently, logical approaches using knowledge representation techniques (e.g. [SPR07]) and combinations of these two kinds of methods have been developed (e.g. [FSPR09], [ARS09]).

Contribution. As far as we know, the methods used for solving author linkage problems are essentially based on name comparisons. SudocAD has a lot of original facets since it uses:

- **Combination of numerical and logical** computations. Numerical computations are used for comparing low level attributes, and logical rules are applied for higher level attributes. The semantic web languages RDFS and OWL, as well as knowledge representation techniques (cf. [CM09]), have been used for developing the system.
- **Super-authorities**. A notion of super-authority is introduced, which is an authority enriched by information present in the whole bibliographic database. Furthermore, if the document database considered does not contain authority records, the super-authority building technique can be used to build such authority records.

⋆ This work benefited from the support of ANR, the French Research National Agency (ANR-12-CORD-0012)

– **Qualitative partition**. The result obtained is an **ordered qualitative partition** of the set of candidates. This ordered partition can be used in different ways, thus allowing different strategies for the linkage problem (in automatic mode as well as in decision-aided mode).
– **Genericity**. The method underlying SudocAD is generic and even though it deals with authors it can be used for other authorities as well, e.g. collective entities.

Another key contribution of the present paper is the evaluation method used (and the results of this evaluation).

The paper is organized as follows. In Section 2, the bibliographic database context of the system SudocAD is presented. The hybrid method, combining numerical and logical aspects, underlying SudocAD is presented in Section 3. The methodology and the results of the evaluation of SudocAD are described in Section 4. Cogui, the tool used for implementing SudocAD, is presented in Section 5. The Conclusion and further work end the paper.

2 The Bibliographic Context

In this section, the bibliographic context of SudocAD is briefly described.

2.1 Bibliographic and Authority Notices, Sudoc and IdRef

Sudoc (cf.[suda]) is the national bibliographic infrastructure for French higher education developed and maintained by ABES (National Agency in charge of Sudoc cf.[abe]). As this infrastructure, Sudoc is both:

– a shared database where bibliographic records are created once, but used by all ;
– a network of professional catalogers, with common tools, and guidelines for using the tools.

The core function of this collective endeavour is to create and maintain database records that describe the documents held or licensed by French Universities and other higher education institutions. Sudoc contains more than ten million of such records (2012 figure). Documents described by Sudoc are mainly electronic or printed books and journals, but also manuscripts, pictures, maps, etc.

A Sudoc record consists of three kinds of information:

– Meta-information
– Descriptive information
– Access points

Meta-information is information about the record itself. It is beyond the scope of this paper.Descriptive information is mere transcription of information that is found in the document to be catalogued. For instance, the descriptive field $Title$ contains the text string that on the title page. The same is true for the descriptive field $Author$, the record has to keep the author's name as it is found in the document, even if the librarian knows that the title page misspelled this name. This strictly descriptive approach aims to identify the publication without any ambiguity. The transcribed information has to be sufficient to distinguish two editions of the same work. But this descriptive approach may make the document harder to find. If the title page, hence the bibliographic record, assumes that the author's name is 'Jean Dupond' whereas the actual name is 'Jean Dupont', library users who are only aware of the actual name will fail to find and access the record in the catalog and then to find and access the document in the library.To overcome this kind of problem, the cataloguing rules prescribe to add the actual name in the record, not instead of but rather besides the one found on the title page. This kind of additional non-descriptive information is called "access point".

Access points constitute the third kind of information that is to be found in a bibliographic record. An access point is a piece of information that focuses on some aspect of the described document that may be relevant for finding the record in a database and that is not directly observable in the document itself. As we have seen above, it can be the actual author's name. It can also be the author's name written according to a specific convention (e.g. 'Dupont, Jean' or 'Dupont, Jean (1956-)').But an access point is not necessarily an alternative textual form of a piece of information that was observed on the described object. It can be the result of the analysis of the document content. For instance, an access point can be a keyword expressing the subject of the document (e.g. 'cars').

Any bibliographic record is the result of the description of the document and the selection of relevant access points. But when is a bibliographic record deemed to be achieved? Theoretically, the description could last forever but cataloguing guidelines set a finite list in which document properties must or may be described. But regarding access points, when to stop? If access points are available to help the user find the record, there can be plenty of them. Should the record describing the book by Jean Dupont and about cars have access points mentioning all variants of Jean Dupont's name and as many access points as there are synonyms for 'cars'? And should this be the case for each book by Jean Dupont or about cars?

In order not to repeat all the variants of a name or a concept in all bibliographic records using this name or this concept as an access point, these variants are grouped in specific records, namely authority records (cf. [idr]). For instance, the authority record for Jean Dupont will contain his name variants. One of these variants will be distinguished as the preferred form. Some guidelines will help librarians choose the preferred form. Some guidelines expect the preferred form to be unique. The preferred form is the only form to be used as an access point in the bibliographic records. If it is unique, it works like an identifier for the

authority record that contains the rest of the variants: the preferred form used as an access point links the bibliographic record to the authority record.

In many recent catalogs, the link from the bibliographic record to the authority record is not the preferred form but the authority number. This is the case in Sudoc. Sudoc bibliographic record access points are just authority record identifiers. When the record is displayed or exported, the Sudoc system follows the link and substitutes a name or a term for the identifier. Sudoc has more than ten million bibliographic records and two million authority records. An authority record is not an entry in a biographical dictionary. It is supposed to contain nothing but information sufficient to distinguish one person from another one described by another authority record. Authority records for people mainly contain information about their names, dates and country (cf. [idr]). Authority records for concepts contain information about their labels and relationships to other concepts (broader, narrower, etc).

2.2 Semantization of Bibliographic Metadata

Bibliographic and authority records have been expressed in RDF. The RDF vocabulary has to meet some expectations:

- to be able to express precise bibliographic assertions,
- to be minimally stable and maintained by a community.

The FRBRoo vocabulary, which is an object-oriented version of FRBR (cf. [FRBb]), meets these expectations:

- It has been developed for fine-grained bibliographic requirements,
- It is well documented,
- It is maintained and still developed by an active community,
- It is connected to other major modeling initiatives.

Firstly, FRBRoo takes its main concepts from FRBR, a prominent model that was developed by the International Federation of Library Associations during the 1990s. FRBRoo keeps core FRBR concepts but claims to overcome some of its alleged limitations. Secondly, FRBRoo is built as an extension of another model for cultural objects, namely CIDOC CRM. CIDOC CRM's main scope is material cultural heritage, as curated by art or natural history museums. It is focused on expressing the various events that constitute an object life, before or after its accession to the museum. FRBRoo imports many of its classes and properties from CIDOC CRM, but needs to forge a lot of new ones, in order to cope with more abstract entities such as texts and works.

The native Sudoc catalog is in UNIMARC. For SudocAD's needs, it was not necessary to convert all UNIMARC fields in FRBRoo. But even for conversion of the fields needed for the reasoning, we had to extend FRBRoo and forge some new classes and properties. The whole ontology used in SudocAD, which is composed of a domain ontology and a linkage ontology, contains a hierarchy of (313) concepts and a hierarchy of (1179) relations. It is presented in the Examples part on Cogui's website (cf. [cog]). A preliminary report about the system SudocAD has been published, in French, on the ABES website (cf. [sudb]).

2.3 SudocAD

In a very general setting, the input and results of SudocAD can be stated as follows:

- **The data**. The input of the system consists of the sudoc catalog [suda] and the person authority base IdRef [idr], as well as a part of the Persee bibliographic database [per]. The part of Persee considered for SudocAD evaluation contains bibliographic records of papers in social science journals.
- **The problem**. Given a Persee bibliographic record d, link reference E to an author of a the paper described by d to an authority A in IdRef, if E and A refer to the same author.
- **The result**. The result is an ordered list of seven pairwise subsets of IdRef: $S(trong)$, $M(edium)$, $W(eak)$, $P(oor)$, $N(eutral)$, $U(nrelated)$, $I(mpossible)$. The order is a qualitative ordering of authorities which could be linked to E. $S(trong)$ contains authorities for which there is strong evidence that they refer to the same author as E. For $M(edium)$, the evidence is less strong etc., and $I(mpossible)$ contains authorities for which there is strong evidence that they do not refer to the same author as E. This result can be used in an automatic mode, the system makes a decision to link or not link r, or in an aided-decision mode, the system presents the ordered list to a person who has to make a decision.

3 The Method

3.1 Principles

The important tasks of the implemented method can be briefly stated as follows (they are detailed in the forthcoming subsections).

– **Linkage knowledge**: comparison criteria and rules.
A preliminary fundamental task consists of building linkage knowledge whose main components are comparison criteria and logical rules. An *elementary comparison criterion* is a function c_δ that assigns to (E, A), where E is a referent to an author in a bibliographic record d and A is a referent to a person authority, a qualitative value representing the similarity of$E$ and A with respect to δ. The elementary comparison criteria built for SudocAD are: c_{denom} (dealing with denominations of authors), c_{dom} (dealing with scientific domains of documents), c_{date} (dealing with date information), and c_{lang} (dealing with the languages in which the documents are written). For instance, for computing $c_{date}(E, A)$, knowledge such as "if the publication date of d is t then E cannot be identical to a person authority A whose birth date is posterior to t."
The set of values of a comparison criterion c_δ is a totally ordered set of qualitative values representing the similarity between E and A with respect to δ (typically: $similar$, $intermediate$, $weak$, $dissimilar$). See Section 3.4 for details.
The (global) comparison between E and A is also expressed as a qualitative value, $S(trong)$ or $M(edium)$ or $W(eak)$ or $P(oor)$ or $N(eutral)$ or $U(nrelated)$ or $I(mpossible)$), which is the conclusion of a logical rule. These logical rules are as follows: If $c_{denom}(E, A) = H_1$ and $c_{dom}(E, A) = H_2$ and $c_{date}(E, A) = H_3$ and $c_{lang}(E, A) = H_4$ then $linkage(E, A) = C$. (cf. Section 3.5).
The notions needed for expressing these rules are gathered in the linkage ontology $\mathcal{O}_L$.
Once this linkage knowledge has been built, the input data is processed by three successive steps as follows.
– **Working Knowledge Base** The whole bibliographic database (sudoc catalog + IdRef) is very large. In the first processing step, it is restricted to a knowledge base $\mathcal{W}$ expressed with the formal ontology (composed of the domain ontology $\mathcal{O}_B$, based on FRBRoo, and on the linkage ontology $\mathcal{O}_L$). $\mathcal{W}$ should have two main properties: it should contain all authorities which may be linked with E in d, and these authorities are called authority candidates. The second property it that $\mathcal{W}$ should be small enough to efficiently perform the computations needed by the linkage problems. See Section 3.2 for the construction of $\mathcal{W}$.
– **Authority Enrichment** In the second step, each authority candidate in $\mathcal{W}$ is enriched, and the result of such enrichment is called a *super-authority*. The links between document records and authority records are used to build these super-authorities. For instance, one can add an attribute 'area of competence' of an authority A and if A is the author of many documents dealing with medicine it is probably relevant to add that medicine is within the area of competence of A. Instead of adding new attributes, it is also possible to specialize information already existing in an authority record. These super-authorities are compared to the information about E, i.e. the entity to be linked, obtained from d. See Sectioncf. 3.3) for more details.
– **Linkage Computations** The linkage task itself is decomposed into two subtasks as follows. For each couple (E, A) and for each comparison criterion c_δ, the value $c_\delta(E, A)$ is computed. This computation uses numerical computations but the resulting value $c_\delta(E, A)$ is a qualitative value (e.g. $similar$ or $intermediate$ or $weak$ or $dissimilar$). These qualitative values are used as hypotheses in logical rules (see Section 3.5) .
Finally, logical rules, having values of comparison criteria (between E and A) as conditions and a value of the global qualitative comparison criterion, $link$, as conclusion are fired. The result is a partition of the authority candidates ordered by decreasing relevance with respect to the possibility of linking E and A. For instance, if $link(E, A) = Strong$, there is strong evidence that E and A can be linked, i.e. that they represent the same entity in the exterior world.

3.2 Working Base and authority candidates

The construction of the working base $\mathcal{W}$ is detailed in this section. The main steps are as follows.

1. For each author name in d, the first task is to represent an author name in d in the same way, say $name$, as denominations are represented in authority records (this may need an alignment between 'author name' in d and 'denomination' in authority records).
2. A set $\mathcal{A} = sim(name)$ of authorities having a denomination close to $name$ is computed. Note that the variants of the name present in the authority record are used for this computation.
3. For each authority A in $\mathcal{A}$, the set $Bib(A)$ of bibliographic records having A as an author or as a contributor with a significant role is computed.
4. The working base $\mathcal{W}$ is obtained by making the union of the authority records in $\mathcal{A}$ and the document records $Bib(A)$ for all A in $\mathcal{A}$.

A fundamental condition is that the function sim is sufficiently robust to author name variations so that if there is an authority in the authority base corresponding to the author whose name is $name$ in d, then this authority is (almost surely) in $\mathcal{A}$. Another reason for considering, in this step, a generous similarity function is that $\mathcal{W}$ should contain sufficient contextual knowledge concerning the authorities in order to remove the ambiguities, i.e. to solve the linkage problem.

3.3 Super-authority

An authority record does not contain a lot of information. Indeed, an authority record for people is not a biographical record, its only goal is to distinguish one person from another one (also described by an authority record). Authority records for people essentially contain information about their names, dates and country (cf. [idr]). Enriching an authority record with information concerning this authority that can be obtained by searching the bibliographic records is a key idea of SudocAD.

In d, for a bibliographic record of a paper in a scientific journal, one has only, besides the names of the authors, the publication date, paper title, language in which the paper is written and a list of scientific domains. For each of these notions, one can possibly enrich an authority record by using information in the document records, $Bib(A)$, in which A is a contributor. For instance, aggregating all domains of records in $Bib(A)$ is generally more precise than a piece of information concerning the competence in A. In the same way, one can compute, and then assign to A, an interval of publication dates. Note that, due to the nature of d (scientific paper), the kinds of contributor considered are restricted to those having a scientific role, e.g. author, PhD supervisor, scientific editor, preface writer. Note that, as in a bibliographic record d within the bibliographic base Persee, there is only, besides the names of the authors, the publication date, , paper title and , language. Thus, super-authorities deal only with information that can be compared with information in d, i.e.,, domain, date, language. The next section stipulates how this new information is computed and defines the comparison criteria.

Remark Note that if a document base does not contain authority records but only bibliographic records then the method for building super-authorities can be used, starting with an authority record containing only a name attribute, to build authority records.

3.4 Comparison Criteria

Due to the nature of Persee bibliographic records, the only information taken into account in SudocAD is: $denomination$, $domain$, $date$, $language$. The corresponding elementary comparison criteria and how their values are computed for a given couple (E, A) are described in this section.

Denomination An author authority record (in the Sudoc base) contains all known variants of an author name, thus the attribute denomination is not used for computing a super-authority.

The name n_d of an author E in d, and a denomination n_A of an authority A, are split into two strings, respectively (n, p) and (n', p'). The first string is the most discriminant part of the name, in our case it corresponds to the family name, and the second string (less important) is composed of first names or first name initials. The strings n, n', p, p' are normalized (transformation of uppercases into lowercases, deletion of accents and redundant spaces, etc.). Two functions, denoted c and c', respectively compare n, n' and p, p'. For both, their result is one of the following qualitative values $I(dentical)$, $S(trong)$, $C(ompatible)$, $D(istant)$, $Dif(erent)$. c and c' use the same distance algorithm (Levenshtein's algorithm) with the same threshold, but due to their different nature (e.g. possibly initials in p's) the two functions are rather different. For instance, they differently use the prefix notion (these functions are completely described on Cogui's website).

The results of these two functions are aggregated as follows, where the qualitative values of $c_{denom}(E, A)$ are coded as follows: +++ stands for *same denomination*, ++ for *close denomination*, + for *distant denomination*, - for *dissimilar denomination*.

```
if (c(n, n') = I or S) then
  if c'(p,p')= I or S return +++;
  if c'(p,p')= C or D return ++;
  otherwise return +);
if (c(n, n') = C) then
  if c'(p,p')= I or S or C return ++;
  if c'(p,p')=  D return +;
  otherwise return -);
if (c(n, n') = D) then
  if c'(p,p')= I or S or C or D return +;
  otherwise return -);
otherwise return -
```

Finally, since an authority A may have a set of denominations, $Denom(A)$, the value of the denomination criterion $c_{denom}(E, A)$ is equal to the maximum value over the set of denominations of A, i.e. if n_d is the name of E,

$c_{denom}(E, A) = max\{c_{denom}(n_d, n_A) | n_A \in Denom(A)\}$.

Domain For each authority candidate A, a domain profile, which is a set of weighted domains $\{(d_1, p_1), \ldots, (d_k, p_k)\}$, is computed as follows.

A bibliographic record has a domain attribute which is a multi-set of domains (the same domain can occur several times). To a bibliographic record B which is in $Bib(A)$, i.e., for which A is a scientific contributor, is assigned a weighted set $\{(d_1, q_1), \ldots, (d_m, q_m)\}$, where $\{d_1, \ldots, d_m\}$ is the set of domains in B and $q_i = 1/\#d_i$ where $\#d_i$ is the number of occurrences of d_i in the domain attribute of B. This set is considered as a vector on the set of domains and the domain profile of A is the sum of these vectors for all documents in $Bib(A)$. Note that this domain profile is a new piece of information in the super-authority of A.

In the same way, a domain profile can also be assigned to document d, with the weight of a domain of d being equal to $1/\#dd$, where $\#dd$ is the number of domains associated with d. The domains associated with d is the domains of the Journal

in which d has been published. We compare the set of domains of the Journal in which d has been published with the weighted list of domains in the authority A.

The similarity measure between two domain profiles $P = \{(d_1, p_1), \ldots, (d_m, p_m)\}$ and $P' = \{(d'_1, p'_1), \ldots, (d'_n, p'_n)\}$ that have been used in SudocAD is

$$\sigma(P, P') = \sum_{i=1}^{m} min(p_i, \sum_{j=1}^{n} p'_j \sigma(d_i, d'_j))$$

In this formula,

$$\sigma(d, d') \in [0, 1]$$

is a similarity between domains. $\sigma(d, d') = 1$ means that $d = d'$ or that d and d' are synonyms and $\sigma(d, d') = 0$ means that it is quite impossible that a given person has a scientific role in a document about d and a document about d'.

The qualitative values of the *domain* criterion are coded as follows : +++ stands for *strong correspondence*, ++ for *intermediate correspondence*, + for *weak correspondence*, - for *without correspondence*, and ? for *unknown* (Bibliographic records do not always have domain attributes, thus it is possible that domain profiles cannot be computed). The qualitative values for the domain criterion are defined as follows, $c_{dom}(d, A)$ is equal to:

- +++ whenever $0.8 < \sigma(P, P') \leq 1$,
- ++ whenever $0.5 < \sigma(P, P') \leq 0.8$,
- + whenever $0.2 < \sigma(P, P') \leq 0.5$,
- - whenever $0 \geq \sigma(P, P') \leq 0.2$.

Note that $c_{dom}(E, A) = c_{dom}(d, A)$ for any author in d since we only use d to compute the domain profile of an author of d.

Date Two notions are used to define the date criterion. The first one expresses compatibility between the publication date p_d of d and the interval of publication dates ($beginPeriod$, $endPeriod$) of A and the other expresses compatibility between the publication date of d and the life interval ($birthDate$, $deathDate$) of A when these dates are known. If only one of these dates exists, the second one is approximated w.r.t. a (reasonable) maximal lifespan.

Three parameters are used: T_1 is the age at which a person can begin to publish , T_2 is the maximal lifespan whenever either the birth date or the death date of an author is unknown and T_3 is used to express that the publication date of d is close to the interval of publication dates. In SudocAD, T_1 is set at 20, T_2t at 100 and T_3 at 10.

The following definitions are used in the table specifying the date criterion.

$inPer = (beginPeriod \leq p_d)$ and $(p_d \leq endPeriod)$;

$closePer =$ not $inPer$ and $(beginPeriod - T_3 \leq p_d)$ and $(p_d \leq endPeriod + T_3)$;

$outPer =$ not $inPer$ and not $closePer$;

$beforeLife = (p_d \leq birthDate + T_1)$;

$inLife = (birthDate + T_1 \leq p_d)$ and $(p_d \leq deathDate))$;

$afterLife = (p_d \geq deathDate)$;

The qualitative values of the *date* criterion are coded as follows : +++ stands for *strong correspondence*, ++ for *intermediate correspondence*, + for *weak correspondence*, - for *without correspondence*, and ? for *unknown*.

The value of the date criterion is given by the following table.

date criterion	$inPer$	$closePer$	$outPer$
$inLife$	+++	++	++
$afterLife$	++	+	+
no LifeDates	++	+	+
$beforeLife$	-	-	-

Language For our experiment, the language is not discriminant and we have chosen a very simple language criterion: if in $Bib(A)$ there is a document written in the same language as d then the value of $c_{lang}(d, A)$ is + and otherwise it is -. Note that, as for the domain criterion, the language criterion actually deals with d and is transferred to each author E in d.

3.5 Linkage

Principles Let us recall some notations. E is an author in d and $\mathcal{A}$ is the set of authority candidates obtained in the working base computation step (see Section3.2). For each (E, A), A in $\mathcal{A}$, and each criterion c_δ, the value $c_\delta(E, A)$ is computed as explained in Section 3.4. These values are used for computing the possibility of linkage between E and A. This possibility is expressed by a (global) comparison criterion called $linkage$. The value of $linkage(E, A)$ is a value in the ordered set: $S(trong)$, $M(edium)$, $W(eak)$,$P(oor)$, $N(eutral)$, $U(nrelated)$, $I(mpossible)$s.

$linkage(E, A)$ is obtained as the conclusion of logical rules whose premises are the values of $c_{denom}(E, A)$, $c_{date}(E, A)$, $c_{dom}(E, A)$ and $c_{lang}(E, A)$. Here is an example of a rule (noted $LS2$ in the table in Section 3.5).

```
If $c_{denom}(E,A) = +++$ and $c_{date}(E,A)=++$ and
   $c_{dom}(E,A)=+++$ and $c_{lang}(E,A)=+$
then $linkage(E,A)=S$.
```

This rule is used as follows. If for a given pair (E, A) all premises are true, then the system concludes that there is strong evidence that E and A represent the same entity. Said otherwise, in this situation, the system cannot distinguish E from A. At the other end, if $linkage(E, A) = I$, this means that the system considers that there is strong evidence that E and A do not refer to the same author.

The values of $linkage(E, A)$ partition $\mathcal{A}$, a class being composed of all A in$\mathcal{A}$ having the same value $linkage(E, A)$. This partition is ordered by decreasing relevance with respect to the possibility of linking E and A.

This partition can be used in different ways, and the choices made in SudocAD for an automatic linkage mode and for a decision-aided mode are presented in Section 4.

In SudocAD The 22 rules used in SudocAD are listed in the following table in which the joker * stands for any value of a criterion as well as the absence of value.

These rules are fired in a specific order and $linkage(E, A)$ is the first value obtained, i.e. as soon as a rule is fired for a pair (E, A) the others are not fired. The chosen order is as follows: LI1, LI2, LU3, LP4, LS1, LS2, LM1, LM2, LM3, LM4, LM5, LW1, LW2, LW3, LW4, LW5, LP1, LP2, LP3, LU1, LU2, Another case.

The order on the set of values of a comparison criterion is also used. If the value of a criterion is positive, say p, then it is assumed that it also has all the positive values $p' \leq p$. The way the rules are fired ensures that the result obtained is equivalent to the result given by 300 rules fired in any order, i.e. used in a declarative way.

Rule	Denom	Date	Dom	Lang	Linkage
LS1	+++	+++	++	+	S
LS2	+++	++	+++	+	S
LM1	+++	*	+++	*	M
LM2	+++	+	++	+	M
LM3	++	+++	+++	*	M
LM4	++	++	++	+	M
LM5	+++	++	+	+	M
LW1	++	++	+	*	W
LW2	++	+	++	+	W
LW3	+	+++	+++	*	W
LW4	++	+	+	*	W
LW5	+++	*	++	+	W
LP1	+++	*	-	*	P
LP2	++	*	*	*	P
LP3	+	++	*	*	P
LP4	*	*	*	-	P
LU1	+	*	*	-	U
LU2	*	+	-	*	U
LU3	*	-	*	*	U
LI1	*	-	-	*	I
LI2	-	*	*	*	I
Other case	*	*	*	*	N

4 Evaluation

4.1 Methodology

It seems difficult to assess the quality of the partition obtained by SudocAD and even to define the quality of such a partition. Thus, we consider two ways of using this partition which can be evaluated by human experts: an automatic mode and a decision-aided mode. In the automatic mode, the system either proposes an authority A to be linked to E or proposes no link. In the decision-aided mode, the system proposes a list of authorities in a decreasing relevance order until the operator chooses one authority or stops using the system.

From the database Persee (cf. [per]), 150 bibliographic records, referencing 212 authors, were chosen at random. For these records, professional librarians had to do their usual work, that is to try to link authors in Persee records to authorities in IdRef [idr] in their usual work environment. That is to say, librarians had the Persee records and usual on-line access to the sudoc catalog and to the IdRef authorities. Librarians had also a limited time, no more than 5 min for linking an author, and they also had to respect the usual constraint to avoid creating erroneous links. For an author E in a Persee record, a librarian could make one of the following decisions:

- link with certainty E to an authority A
- link with uncertainty E to an authority A and suggest other possible authorities
- refrain with certainty for linking
- refrain with uncertainty for linking and suggest possible authorities

Note that the last two situations should normally prompt the librarian to create a new authority record.

An expert librarian analyzed the results of this first step. He was not involved in this step and could use any source of information (e.g. the web) and he had no time limited. The results obtained after this step are hoped to be the best possible linkages in such a way that the comparison of these results with those obtained by SudocAD is meaningful.

Expert Results	Number
Link with certainty	146
Link with uncertainty and other choices	3
Link with uncertainty and no other choices	19
No link with certainty	37
No link without certainty and other choices	7
No link without certainty and no other choices	0

4.2 Automatic Linkage

We considered four different ways of automatic linkage, listed as follows from the most restrictive to the least restrictive linkage.

- AL_1: If the best class, i.e. $S(trong)$, contains only one authority, then E is linked to this authority;
- AL_2: If the union of the two best classes, i.e. $S(trong)$ and $M(edium)$, contains only one authority, then E is linked to this authority;
- AL_3: If the union of the three best classes, i.e. $S(trong)e$ and $M(edium)$ and $W(eak)$, contains only one authority, then E is linked to this authority;
- AL_4: If the union of the four best classes, i.e. $S(trong)$ and $M(edium)$ and $W(eak)$ and $P(oor)$, contains only one authority, then E is linked to this authority.

For the comparison between the expert choice and one of these methods we considered that the answer given by the method corresponded to one of the following decisions:

- *Good decision:* either when the expert links with certainty E to A and the method links E to A or when the expert does not link with certainty E and the method does not propose a link;
- *Acceptable decision:* either when the expert links with uncertainty E to A and the method links E to A or when the expert does not link with certainty E and the method proposes a link to a possible candidate proposed by the expert;
- *Bad decision:* either when the expert links with certainty E to A and the method links E to $A' \neq A$ or when the expert does not link with certainty E and the method proposes a link;
- *Prudent decision:* when the expert links with or without certainty and the method does not propose a link.

The means of these parameters for the 212 authors occurring in the 150 Persee bibliographic records are as follows.

Method	Good	Acceptable	Bad	Prudent
AL_1	54.7%	0%	1.89%	43.4%
AL_2	77.36%	0.47%	1.89%	20.28%
AL_3	80.19%	0.47%	3.77%	15.57%
AL_4	86.79%	0.94%	6.6%	5.66%

The results were very positive. The choice of a method AL_i could be guided by the importance given to the quality of the decision made by the system. If bad decisions have to be strictly avoided while having a significant number of good decisions, then AL_2 is a good choice. If the number of good or acceptable decisions is the main criterion, then AL_4 can be chosen. Note furthermore, that most bad decisions of any method AL_i arise from errors in the bibliographic database and are not related to intrinsic flaws in the system itself. Indeed, when there are false links from bibliographic records to authority records, these erroneous links can lead to building incorrect super-authorities.

4.3 Decision-Aided

In a decision-aided mode, the system presents an ordered list of candidate authorities to a human operator. This list is presented in decreasing order of relevance $S(trong)$, $M(edium)$, $W(eak)$, etc. until the operator chooses one authority to be linked to E or stops and concludes that there is no authority which can be linked to E.

Three classical Information Retrieval parameters were considered to evaluate the use of our system in a decision-aided mode. They use the following sets of authorities:

- $Cand$ is the set of authority candidates in the working base,
- $Impos$ is the set of authorities related by $I(mpossible)$ to E,
- $Selec$ is the union of the authority linked to E by the operator and, when there is uncertainty, the set of possible authorities proposed by the operator.

The parameters are then defined as follows.

- *Recall*
  $recall = |Selec \cap (Cand \setminus Impos)|/|Selected|$
- *Precision*
  $precision = |Selec \cap (Cand \setminus Impos)|/|(Cand \setminus Impos)|$
- *Relevance*
  $relevance = |Selec|/MaxPos$,
  where $MaxPos$ is the last position in the ordered result of an authority in $Selected$.

The means of these parameters for the 212 authors occurring in the 150 Persee records and without considering the candidates in $I(mpossible)$ or $U(nrelated)$ are as follows.

Expert choice	Recall	Precision	Relevance
No link with uncertainty	92.86%	43.94%	56.12%
Link with certainty	100%	78.76%	94.32%
Link with uncertainty	98.48%	68.25%	91.71%

Note that when the librarian has chosen to link, the relevance is higher than 90%. This means that when the librarian searches the candidates in the order given by SudocAD, then the first one is, with few exceptions, the good one.

More importantly, note also that these parameters allow only a partial evaluation of the decision-aided mode. Indeed, there are two different situations: the operator can decide either to link or not to link E to an authority. If he decides to link, the recall is not relevant, because in this case the recall is equal to 1 unless $Selected \cap Impossible \neq \emptyset$, which is unlikely. If he decides not to link, *Selected* is empty. It is possible to propose other parameters but for a significant evaluation one should carry out an experiment comparing two systems. This is further work that is planned for a decision-aided use of our system.

5 COGUI

The system SudocAD has been developed on COGUI (see [cog]) and uses web services for accessing an RDF version of the Sudoc catalog and Sudoc authority records. COGUI is a platform for designing, building and reasoning –with logical rules (cf. [BLMS11])– on conceptual graph knowledge bases. COGUI was used at each stage of the project (cf. [CM09]). Information concerning SudocAD can be found on the Cogui website, on the examples page (cf. [exa]).

– *Ontology*The ontology used in SudocAD contains a hierarchy of (313) concepts and a hierarchy of (1179) relations. The root of the linkage relations is *liage : binding vocabulary relation (Resource,Resource)*. A relation has a signature which indicates its arity and the maximal types an attribute can have (e.g. the two attributes of the relation *liage : liageAuthority (Person,Person)* have to be of type (less than or equal to) Person.
– *Input Data* Input data consists of a part of the Sudoc database, which consists of a catalog and an authority base, and of a Persee record. As explained in section 2 the Sudoc database was translated into RDF. A part of it, containing authority records whose denominations are close to the name of an author in the Persee record and the bibliographic records in which these authorities have a scientific role, is accessed through web services and then imported into Cogui through the import RDF/S natural mode tool.
– *Scripts* The method described in section 3 is implemented through scripts which use queries for searching the data graph.

6 Conclusion and further work

As far as we know, SudocAD is an original system whose main facets are: the combination of numerical and logical computations; the notion and construction of super-authorities; the type of the result, a qualitatively ordered partition of the set of candidates, which allows to build different strategies for the linkage problem (in an automatic mode as well as in a decision-aided mode); and the genericity of the method, SudocAD deals with authors but can be used for other authorities as well, e.g. collective entities. Note also that the super-authority step can be used to build authority records for a document database only containing bibliographic records. The results obtained by SudocAD have been evaluated by a rigorous and demanding method and are very positive. ABES plan to develop SudocAD into a system usable in production conditions.

Improvements can be made when using the same methodology. Here are some of them we plan to investigate in the near future:

– **Improve criteria**. For instance, the domain criterion is the only attribute concerning the content of a document. This can be improved by using the title of a document, which gives more precise information concerning the content of a document than the domain attribute. The date criterion can also be improved by considering the time in which data are relevant intervals (e.g. period during which a person is affiliated to an institution or has a particular function).
– **Enrich the super-authority and add new criteria**. A bibliographic or document database is a rich resource that can be more intensively scanned. For instance, neither co-authors nor the affiliation relationships between institutions and authors have been used.
– **Narrow the search results**. The genericity of the method can allow to iteratively use it in order to reinforce the results. First, use a methodology similar to SudocAd for linking institutions, use these links for linking authors (using the affiliation of authors to institutions) then check or modify the links between institutions.

The algorithms used in SudocAD assume that the bases are correct. Each record, either bibliographic record or authority record, can indeed be assumed to be correct (especially when they have been built by librarians and not by an automatic system). However, links between records can be erroneous, especially since some of them have been automatically computed when the considered base has been obtained by automatically merging several bases. Further work are needed to define the quality of document bases, for improving the quality of the links existing in a document base, and for taking into account the quality of the considered bases in linkage problems.

References

[abe] ABES. `http://www.abes.fr/`.

[ARS09] Arvind Arasu, Christopher R, and Dan Suciu. Large-scale deduplication with constraints using dedupalog. In *in: Proceedings of the 25th International Conference on Data Engineering (ICDE*, pages 952–963, 2009.

[BGMM+09] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Qi Su, S. Euijong Whang, and J. Widom. Swoosh: a generic approach to entity resolution. *The VLDB Journal*, 18(9-10):255–276, 2009.

[BLMS11] Jean-François Baget, Michel Leclère, Marie-Laure Mugnier, and Eric Salvat. On rules with existential variables: Walking the decidability line. *Artif. Intell.*, 175(9-10):1620–1654, 2011.

[cid] The CIDOC CRM . `http://www.cidoc-crm.org/`.

[CM09] M. Chein and M.-L. Mugnier. *Graph-based Knowledge Representation*. Springer, London, UK, 2009.

[cog] COGUI. `http://www.lirmm.fr/cogui/`.

[dCLGdS12] M.G. de Carvalho, A.H.F. Laender, M.A. Goncalves, and A.S. da Silva. Genetic programming approach to record deduplication. *Knowledge and Data Engineering, IEEE Transactions on*, 24:3:399 – 412, 2012.

[EIV07] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. Duplicate record detection: A survey. *TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, page 2007, 2007.

[exa] COGUI examples. `http://www.lirmm.fr/cogui/examples.php#sudocad`.

[frba] Functional Requirements for Bibliographic Records . `http://www.ifla.org/publications/functional-requirements-for-bibliographic-records`.

[FRBb] Object Formulation of FRBR . `http://www.cidoc-crm.org/frbr_inro.html`.

[FS69] I.P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 1969.

[FSPR09] F. Fatiha Sais, N. Pernelle, , and M.-C. Rousset. Combining a logical and a numerical method for reference reconciliation. *Journal of Data Semantics*, pages 66–94, 2009.

[GBVR03] L. Gu, R. Baxter, D. Vickers, and C. Rainsford. Record linkage: current practice and future directions. Technical Report 03/83, CSIRO Mathematical and Information Sciences, 2003.

[Gom02] Shanti Gomatam. An empirical comparison of record linkage procedures. *Statist. Med.*, 21(1):1485–1496, 2002.

[HS98] M.A. Hernndez and S.J. Stolfo. Real-world data is dirty: data cleansing and the merge/purge problem. *Data Min. Knowl. Discov.*, 20(2(1)):9–37, 1998.

[idr] IdRef:authority files of the Sudoc database. `http://en.abes.fr/Other-services/IdRef`.

[MAS12] Suchanek F. M., S. Abiteboul, and P. Senellart. Paris: Probabilistic alignment of relations, instances, and schema. In *Proceedings of the VLDB Endowment, Vol. 5, No. 3*, pages 157–168, 2012.

[NKAJ59] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James. Automatic linkage of vital records. *Science*, 1959.

[NRST09] N.R. Neil R. Smalheiser and V.I. Torvik. Author name disambiguation. *Annual Review of Information Science and Technology (ARIST)*, 43, 2009.

[per] PerseeD. `http://www.persee.fr/web/guest/home`.

[rda] RDA: Resource Description and Access. `http://www.rda-jsc.org/rda.html`.

[SD05] P. Singla and P. Domingos. Object identification with attribute-mediated dependences. In *Proc. of PKDD 2005*, pages 297–308, 2005.

[SE13] Pavel Shvaiko and Jérôme Euzenat. Ontology matching: State of the art and future challenges. *IEEE Trans. Knowl. Data Eng.*, 25(1):158–176, 2013.

[SPR07] F. Sais, N. Pernelle, , and M.-C. Rousset. L2r: a logical method for reference reconciliation. In *Proc. of AAAI 2007*, pages 329–334, 2007.

[suda] SUDOC. `http://www.abes.fr/Sudoc/Sudoc-public`.

[sudb] sudocAD. `http://www.abes.fr/Sudoc/Projets-en-cours/SudocAD`.

[W.E06] Winkler W.E. Overview of record linkage and current research directions. Technical report, U.S. Census Bureau, 2006.

[W.E08] Winkler W.E. Record linkage references. Technical report, U.S. Census Bureau, 2008.