Synthesis Lectures on
Information Concepts, Retrieval, and Services

SYNTHESIS
COLLECTION OF TECHNOLOGY

Reagan W. Moore

Trustworthy Communications and Complete Genealogies

Unifying Ancestries for a Genealogical History of the Modern World

Springer

Synthesis Lectures on Information Concepts, Retrieval, and Services

Series Editor

Gary Marchionini, School of Information and Library Science, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

This series publishes short books on topics pertaining to information science and applications of technology to information discovery, production, distribution, and management. Potential topics include: data models, indexing theory and algorithms, classification, information architecture, information economics, privacy and identity, scholarly communication, bibliometrics and webometrics, personal information management, human information behavior, digital libraries, archives and preservation, cultural informatics, information retrieval evaluation, data fusion, relevance feedback, recommendation systems, question answering, natural language processing for retrieval, text summarization, multimedia retrieval, multilingual retrieval, and exploratory search.

Reagan W. Moore

Trustworthy Communications and Complete Genealogies

Unifying Ancestries for a Genealogical History of the Modern World

Third Edition

Springer

Reagan W. Moore
School of Information and Library Science
University of North Carolina
Chapel Hill, NC, USA

ISSN 1947-945X ISSN 1947-9468 (electronic)
Synthesis Lectures on Information Concepts, Retrieval, and Services
ISBN 978-3-031-65225-7 ISBN 978-3-031-65226-4 (eBook)
https://doi.org/10.1007/978-3-031-65226-4

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The 3rd edition of *Trustworthy Communications and Complete Genealogies: Unifying Ancestries for a Genealogical History of the Modern World* analyzes the properties of a Unifying Ancestry for Western Europe. A genealogy that connects to the Unifying Ancestry for a national community is complete. However, all members of the national community should be able to connect. The ancestors of any member could be chosen as the Unifying Ancestry. A Coherence metric is proposed that differentiates between possible Unifying Ancestries and selects the best one. The Unifying Ancestry is complete when it connects to an Age of Progenitors when any person alive is a common ancestor of all members of the National Community. The Age of Progenitors is complete when the number of generations needed to trace ancestry back to the Age of Progenitors is less than an analytic estimate. The smallest Unifying Ancestry that matches the analytic estimate has the optimal size.

An extended version of the Research Genealogy based on 348,844 persons is used. The best Unifying Ancestry for Western Europeans is based on 38,533 common ancestors of the Kings of Spain, Great Britain, Belgium, Denmark, the Netherlands, Norway, and Sweden. These lineages are highly intermarried, link to the noble houses of Europe, and extend back more than 50 generations.

A complete genealogy has three main sections: the immediate ancestors of a root person; connections to the Unifying Ancestry for their national community; and connections from the Unifying Ancestry to an Age of Progenitors, when all persons are ancestors of the current members of the national community. This approach requires that a complete Unifying Ancestry be created. An example is provided.

Example lineages of Maximal Ascent to Charlemagne are provided for 27 persons identified by both Hart *The 100: A Ranking of the Most Influential Persons in History* (Hart 2000) and Palmer *A History of the Modern World* (Palmer & Colton 1952).

Chapel Hill, NC, USA — Reagan W. Moore

References

Hart, M. (2000). The 100: A ranking of the most influential persons in history. Citadel.
Palmer, R. R., & Colton, J. (1952). A history of the modern world. Alfred A Knopf.

Contents

Introduction

Information Science provides a way to identify properties of genealogies, construct a Unifying Ancestry for a national community, and link the Unifying Ancestry to a Genealogical History of the Modern World. The challenges that need to be addressed include the extraction of information from multiple sources, development of algorithms that analyze genealogy properties, and demonstration that genealogies can be linked to historical events. We start with an analysis of the properties that are needed to trust information derived from multiple sources.

Communication is the exchange of information between a sender and a receiver. Communication is trustworthy if both the sender and receiver interpret the information within the communication using the same context. Context is defined by a knowledge base that organizes relationships between the information elements within the communications. If both participants share the same context for interpreting the information, the communication will succeed. If they do not share the same context, the interpretation by the receiver will differ from the intent of the sender, and the communication will fail. If the communications lack essential information, no useful conclusions can be drawn from the information exchange.

The communication context includes:

- the date that the communication was initiated,
- the source of the information contained within the communication,
- the structural relationships needed to parse information from the communication,
- the semantics used to interpret information elements, and
- a set of knowledge relationships that are defined between the information elements.

R. W. Moore, *Trustworthy Communications and Complete Genealogies*, Synthesis Lectures on Information Concepts, Retrieval, and Services,
https://doi.org/10.1007/978-3-031-65226-4_1

Each participant has an internal knowledge base that they use to organize information for exchange. In addition, there usually is a community-consensus knowledge base that governs terms used in the domain of discourse. When both participants agree on a shared context and reference the same community-consensus knowledge base, information can be reliably exchanged.

How can communication be trustworthy when a shared context is not available? Two worthwhile cases are communication with the future, and interpretation of communication from the past. In both cases, a shared context may not be available. These two cases can be used to quantify properties that trustworthy communications should possess.

Communication with the Future

Preservation is communication with the future. Archivists preserve records that document historical events for access by future generations. The Preservation community has developed two international standards that describe the communication context:

- ISO 14721—the Open Archival Information System (Consultative Committee for Space Data, 2012). OAIS defines a context for each record composed of provenance information, authenticity information, integrity information, description information, representation information, and identification of a knowledge community.
- ISO 16363:2012—Audit and certification of Trustworthy digital repositories. (Technical Committee: ISO/TC 20/SC 13 Space Data, 2012). ISO 16363 defines the information that is needed to track whether the required OAIS information is present, how well the Archives are being managed and whether the Archives are trustworthy.

Can the trustworthiness assessment criteria for Archives be used to define a context for trustworthy communications?

The preservation community effectively defines a context for interpreting the records that comprise a communication with the future. The context consists of:

- Representation information that defines how to parse the record to extract information.
- Provenance information that defines who created the record and who submitted the record to the preservation environment.
- Authenticity information that assigns a unique identifier to the record and tracks whether versions of the record have been created.
- Integrity information that defines whether the record has been corrupted.
- Description information that defines the meaning of the record.
- Specification of a knowledge community that provides an external knowledge base for interpreting relationships between records.
- Management information for tracking changes to the preservation environment.

- Usage information for tracking recipients of the records.
- Security information for tracking unauthorized access.
- Auditing information for tracking the internal operations of the preservation environment.

Given this context, a person in the future should be able to correctly interpret and use information contained in the preserved records. For trustworthy preservation, either the context needed for successful communication is explicitly created and stored with the associated records or procedures are provided for dynamically generating the context. The records together with the context and procedures are archived for use in the future. However, the external knowledge base is not archived. Preservation depends on the continued existence of a knowledge community that maintains the community-consensus knowledge base.

Communication from the Past

The context needed to interpret communication from the past may not be available. Instead, a context may need to be created based upon the information content derived from multiple communications. Effectively, a set of multiple communications is turned into a knowledge base by analyzing relationships between the information elements present within the communications. Relationships that can be established between the information elements can then be used to analyze the trustworthiness of the communications by comparing the derived knowledge base with a community-consensus knowledge base. Since every knowledge base has properties related to Consistency, Correctness, Connectivity, Closure, Completeness, and Coherence, these properties can be used to evaluate the resulting context.

The steps needed to build a missing context for interpreting communications from the past are:

- Identify the information elements that will be extracted from the communications.
- Associate a list of sources with each information element.
- Identify the types of relationships that will be established between the information elements. Examples might be spatial relationships, temporal relationships, semantic relationships, and familial relationships.
- Build a graph database that uses the relationships to define links between the information elements.
- Identify a community-consensus knowledge base for the domain of discourse.

- Analyze the epistemological properties of knowledge bases related to Consistency, Correctness, Connectivity, Closure, Completeness, and Coherence. This will require defining metrics for evaluating each property that are relevant to the communication domain of discourse.
  - Consistency measures whether all the attributes needed to interpret information elements are provided.
  - Correctness measures whether information values fall within acceptable ranges.
  - Closure measures whether information elements are isolated, disconnected from the rest of the information elements extracted from the communications.
  - Connectivity measures how the information elements may be grouped together.
  - Completeness measures whether all information elements can be linked to a unifying topic. The unifying topic describes the intent of multiple communications. The unifying topic must also be linked to a community-consensus knowledge base (through creation of external relationships).
  - Coherence measures whether the derived collection is missing essential information needed to draw conclusions about the unifying topic (through the analysis of internal relationships).
- Finally, construct bi-directional links between the information elements in the database and the information elements in the community-consensus knowledge base. Once this is done, it then becomes possible to interpret the information content present in the communications.

This procedure constructs a context comprised of the relationships between information elements extracted from multiple communications. The multiple communications can be interpreted correctly when both the sender and receiver agree on how the information items are linked to a community-consensus knowledge base.

Genealogy Test Case

Genealogists interpret communications from the past. A practical example of the generation of a local knowledge base is the construction of a genealogy. Records that document historical events are parsed to extract information about marriage, birth, and death; residence, titles, and occupations; familial relationships; and education, religion, and cause of death. Genealogies are viewed as trustworthy if they extract information from authoritative sources. Primary records that document historical events are usually considered to be authoritative. Primary record examples include birth certificates, marriage licenses, tombstones, etc.

Genealogists would like to extract information from authoritative sources, in the expectation that a trustworthy genealogy will then be created. Genealogists rely upon provenance information (sources), authenticity information (the expectation that the

source has not been modified), descriptive information (type of historical events), and representation information (how to interpret dates and locations). The trustworthiness of a genealogy then relies strongly upon the trustworthiness of the sources from which the information has been extracted.

Note there is a strong synergy between the information that Archives require about records and the information that Genealogists parse. A genealogy can be viewed as an index into the records in an archive that identifies familial relationships between the persons involved in the historical events.

The preservation standards for trustworthiness were not published until 2012. Typical sources used for genealogies predate the development of the standards. Also, preservation of records is done independently of the accuracy of the information contained within the records. Genealogists need not only sources that are authoritative, but also sources that contain accurate information. Can metrics be defined that evaluate the accuracy of the information contained within the sources that represent communications from the past?

A genealogy is a knowledge base that links information about persons. Every genealogy has generic properties that include Consistency, Correctness, Closure, Connectivity, Completeness, and Coherence. For genealogies, we can evaluate:

- Consistency—identify the attributes that need to be extracted about each person and verify that each person in the genealogy has the standard set of attributes.
- Correctness—identify external constraints that the attributes should not violate. For example, we can verify that the ages at marriage, birth of a child, and death fall within accepted biological age ranges.
- Closure—verify that each person in the genealogy has a connection to every other person in the genealogy.
- Connectivity—identify the ways in which persons may be grouped, including coalescence of lineages to common ancestors.
- Completeness—identify a connection to a unifying ancestry for all members of a national community. The unifying ancestry should connect to a community consensus knowledge base of historical events.
- Coherence—Identify whether essential information is missing that is needed to identify progenitors.

The Completeness property can be interpreted as the inverse of relationships related to Consistency. In addition to verifying that all the members of a genealogy have a standard attribute such as a familial relationship, a set of progenitors are defined to which all members of the national community should be able to link their ancestry. By linking persons in the genealogy to these progenitors, a genealogy can be immediately integrated with other genealogies that also have lineages to the progenitors. Effectively, a genealogy is inverted from a focus on the ancestors of a root person to a focus on the descendants of a group of progenitors. The Unifying Ancestry is the link between a focus on ancestors of

a root person and a focus on Progenitors that are ancestors of all members of a national community.

The Unifying Ancestry should connect to a community-consensus knowledge base. An example is identifying lineages from the genealogy to historically important persons identified in history books. Such linking assumes that it is possible to identify ancestors for historically notable persons, and then link the genealogy to the same ancestors. By linking the ancestors to historical events, a Genealogical History of the Modern World can be constructed. This, in turn, becomes a community knowledge base for genealogies.

The Coherence property requires that the genealogy be analyzed to determine whether essential information needed to identify progenitors is missing. Does a Unifying Ancestry have the required connectivity to enable all members of a national community to link their ancestry to a set of common progenitors? The coherence metric calculates how many generations into the past lineages need to be traced to find a progenitor. Genealogies that require fewer generations have a better representation of lineage coalescence.

Genealogies serve as a useful test case for the evaluation of collections of communications. Each communication (source) is parsed for information, the information is organized in a database (genealogy), and analyses are then performed upon the database to verify collection properties. Relationships between the information elements are used to construct a local knowledge base, such as the Progenitors of a Unifying Ancestry. The local knowledge base can then be used to analyze the trustworthiness of the genealogy by linking the local knowledge base to external knowledge bases, in this case history books.

An implication is that complete genealogies have three sections: the immediate ancestors of a root person, links to a Unifying Ancestry for their national community, and links from the Unifying Ancestry to the Age of Progenitors. All persons of Western European descent are related through links to a Unifying Ancestry. A genealogy is complete if it connects to the Unifying Ancestry. A Unifying Ancestry is complete if it connects to the Age of Progenitors. The Age of Progenitors is complete if the number of generations to the Age of Progenitors is less than an analytic estimate. This book explores all three forms of completeness.

Unifying Ancestries

Genealogies possess a fundamental symmetry that is driven by two competing processes, the doubling of the number of potential ancestors each generation, and the exponential growth of lineage coalescence. Typically, by the sixth generation, lineages will start to coalesce. Multiple ancestors of the root person of the genealogy will have the same parents. By 1325 AD, the number of potential ancestors each generation exceeds the population of Europe. Lineages must then coalesce because there are not enough persons to fill all the potential ancestors for that generation. The degree of coalescence increases the further the genealogy is traced into the past. When the number of own-cousin lineage

coalescences exceeds the population size, you reach the Age of Progenitors. Before this epoch, all persons are ancestors of the current members of the national community.

The descendants of the progenitors constitute a Unifying Ancestry. A research question is whether this Unifying Ancestry is sufficient to link all members of a national community. Can a set of progenitors be found that comprise the common ancestors of all members of a national community such as persons of Western European descent? One measure of completeness for a genealogy is to demonstrate that a person's ancestors are linked to the Unifying Ancestry for their national community. The Unifying Ancestry represents a knowledge base whose properties can be analyzed and linked to community-consensus knowledge bases such as history books.

Every national community has a Unifying Ancestry to which all members of the national group should be able to link their ancestry. A genealogy can be considered complete when relationships can be defined to all other members of the relevant national community. A Unifying Ancestry simplifies this task, since once you have a connection to your community's Unifying Ancestry, it is the responsibility of the other members of your national community to make their own connections.

A Unifying Ancestry is viable if connections can be made using events within recorded history. This means that a Unifying Ancestry needs to be based on historical fact, rather than being an artefact of cultural tradition. A Unifying Ancestry needs to have a high degree of connectivity through marriages between ancestral lineages. A connection to the Unifying Ancestry should directly lead to familial relationships with the other members of the Unifying Ancestry. Finally, a Unifying Ancestry should contain historically notable ancestors for each national community.

For all genealogies there is a correlation between availability of information about ancestors, and the number of ancestors that can be found. When lineages are traced far enough into the past, typically the only persons that can be identified are historically notable persons such as noble houses. A Unifying Ancestry should include information about historically notable persons who are members of the national community and provide lineages to these historically notable persons. The noble families of Western Europe are a good source for the construction of a Unifying Ancestry. They have the required high degree of intermarriage, represent historically notable people, and extend back 50 generations.

The existence of a Unifying Ancestry is driven by biological, historical, and social factors:

- Biologically, the number of potential ancestors doubles every ancestral generation. By 1325, the potential number of ancestors in a single generation exceeds the available population of Europe. A Unifying Ancestry for Western Europeans could then be created by creating a genealogy based on all the persons alive in 1325 in Europe. Fortunately, we can create a much smaller Unifying Ancestry.

- Historically, the amount of information available about persons decreases as you go back in time. Eventually, the only persons for which we can find information are historically notable persons. This suggests that a Unifying Ancestry will connect to historically important persons, such as noble families.
- Socially, the children of historically notable persons tend to marry the children of historically notable persons. This is particularly true for the Royal Families of Western Europe, the Kings of Belgium, Great Britain, the Netherlands, Denmark, Norway, Spain, and Sweden. A genealogy based on the common ancestors of these Royal Families is a strong candidate for a Unifying Ancestry for persons of Western European descent and will include historical events that comprise a history of the modern world. The lineages for the Royal Families comprise the most heavily researched lineages and therefore will be the most trustworthy. The Royal lineages connect to the noble houses of Western Europe. The Royal Families provide lineage coalescence between national communities, while the noble houses provide lineage coalescence within a national community.

A question that must be addressed is whether the descendants of Kings will include commoners, or whether the descendants will only include nobles. An extensive analysis of the descendants of King George I, Elector of Hanover, was published in "The Book of Kings: A Royal Genealogy". (McNaughton, 1973) McNaughton developed an authoritative genealogy for the descendants through a 20-year exchange of personal communications with the Earl Mountbatten of Burma and Mountbatten's relatives. McNaughton identified 2962 descendants with birth locations in 51 countries, ranging from Europe to North and South America to Eastern Europe, the Middle East, and the Far East. The descendants included titled nobility, clergy, military, doctors, and commoners. The descendants of King George I provide an illustrative example of the number of generations that descents must be traced to find non-royal descendants. Typically, younger siblings become members of the peerage. In the 5th generation from George I, Gustav Karl von Reichenbach-Lessonitz is the Count of Reichenbach-Lessonitz. When younger siblings emigrate to other countries, their children tend to marry commoners. In the 9th generation from King George I, Maximilian von Pagenhardt was born in the United States and married a commoner. The expectation is that lineages linking commoners to Royal Families will require tracing ancestry back at least 7–11 generations.

The design of a Unifying Ancestry (genealogy knowledge base) raises multiple issues:

- How can lineage coalescence be measured?
- How does lineage coalescence lead to a Unifying Ancestry?
- Can a Unifying Ancestry encompass multiple national communities?
- Can a Unifying Ancestry be based on common ancestors of Royal Families?
- How many generations of descendants must be traced from Kings to find non-royal descendants?