The 21st European - Japanese Conference on Information Modelling and Knowledge Bases (EJC2011) 6-10 June 2011, Tallinn, Estonia

Representation and Retrieval of Uncertain Temporal Information in Museum Databases

Miika Nurminen , Anneli Heimbürger
University of Jyväskylä
Faculty of Information Technology
Department of Mathematical Information Technology (MIT)

Outline

1. On Museum Information Systems
2. Representing Temporal Information in *Duo*
3. Proposed Temporal Model
4. Evaluation
5. Conclusion

Museum Information Systems

- Museum information systems (MIS) form a diverse class of collection management and cataloging applications spanning both a multitude of domains (cultural heritage, arts, science, etc) and varying functionality.
- Culture historical information provides a rich and challenging domain for data management, both from temporal and general perspective
  - The number of database tables and metadata fields on a MIS is typically rather large for a cataloging application (i.e. tens of tables each containing multiple fields and relationships), especially compared to library systems.
  - A multitude of complex metadata can be attached to a given object, combining "well-formed" (static, precise, well-known) and ambiguous information
  - Researchers may have unanticipated information needs, ad hoc query -like capabilities are often needed.
- Standards and integration efforts for museum information exist, but in practice the databases used in museums are not interoperable in general.
- A general trend: shifting from item-centric cataloging (physical objects with fixed fields as the primary entities) to event-centric documentation, concentrating on the events (e.g. manufacturing, ownership, documentation, publication) related to the objects.

MIS and Temporal Models for Imperfect Data

- Even though temporal data is an important aspect of museum data, models developed in other disciplines have had relatively little impact on current MIS.
  - While elaborate temporal databases, query languages and logics exist, they do not work well with uncertain temporal information, or do not focus on the problems relevant to information retrieval - especially if precise and uncertain information were to be used with the same interface
  - On the other hand, while a number of techniques to handle imperfect information have been devised both for databases and AI applications, their practical applicability to temporal information is cumbersome from the end user's point of view.
  - Approaches based on temporal granularities (algebraic characterization of years, months, days and other enumerable mappings to the time domain) seem most promising from the MIS point of view, but even that seems too expressive to be easily implemented in a conventional, SQL-based database.
- Flexible, expressive, and easy-to-use –structures that allow incomplete and imprecise information but still support querying and are applicable with relational databases are still needed.
- Our data model can be seen as a limited way to present user-defined symbolic time granularities for anchored temporal primitives supporting indeterminacy and conversions between granularities.

Duo – a Collection Management System

http://users.jyu.fi/~minurmin/duo/

- Duo is the primary system for collection management and museology student projects in JYU Museum
  - Implemented as two-tier client/server database application
  - In use since 2003, includes ca. 37000 items, tens of users, installations in several small museums
  - Provides a framework for other db applications as well (e.g. art database Arte)
- Duo has the typical functionalities for a museum information system
  - Grouping for collections (donations, collection records, and collection objects), exhibition and borrowings management, object placement information, keywords.
  - Varying metadata depending on collection object type (physical object, photograph, recording, book, or newspaper article).
  - Personal data about people related to object documentation.
  - Ad hoc querying, backlink search, image management, html reports.
- What sets Duo apart from most other museum databases is the high level of normalization in the database schema to keep the vocabulary as controlled as possible and minimize typing errors using lookup lists. All related concepts are collected together to ease searchability.

Representation of Temporal Information in Duo

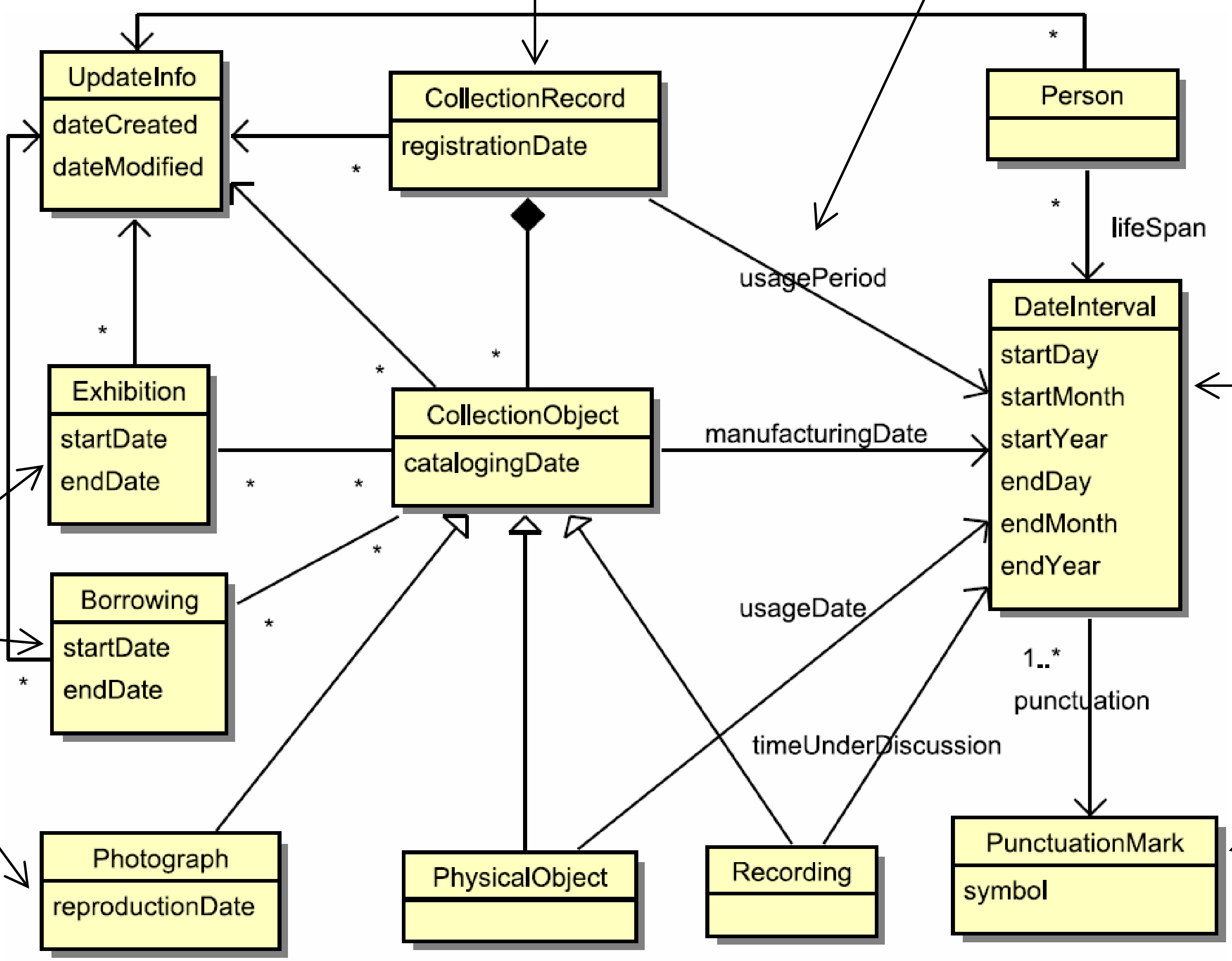
- Standard database DATE type is used for some ”certain” dates (e.g. logging and modification information exhibition dates, check out dates)
- A custom *DateInterval* table is used to represent an uncertain temporal interval.
  - Depending on the metadata field, user can see only years and interval marks. For more specific fields (e.g. lifespan), days and months can be edited as well.
  - Any field can be left empty
- DateInterval uses a *Punctuation mark* between start and end dates.
  - PunctuationMark introduces a number of conventions that can not be easily be reflected in searches (e.g. ”-”:normal interval, ”ca.”:uncertain year, ”decade” to reflect a specific decade symbolically)
  - In practice, the potential semantics in interval mark is not currently accounted for in queries, serves only as descriptive information
- Most of the historical information is uncertain in the first place, and in many cases the researcher is not interested in exact dates, but the more general temporal periods (e.g. 50s, beginning of the century) related to the objects. To support this uncertainty with searchable structure, DateInterval and PunctuationMark tables are used to store most of the collection metadata-related information.

Duo Database Schema for Temporal Entities

Groups objects together

Uncertain dates or intervals

Tracks changes in the database



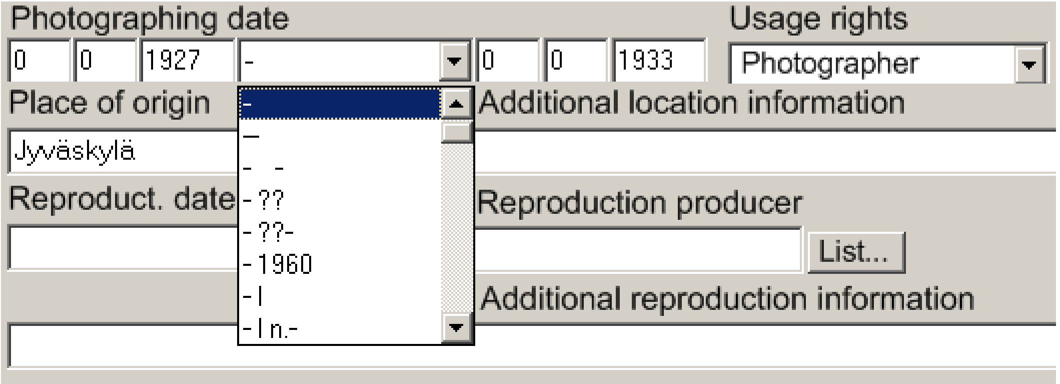
Custom table representing uncertain intervals

Exact (SQL) dates

Additional symbolic information about the interval

Example Form With Uncertain Interval

- Duo screen shot from a data entry form for photographs (translated from Finnish)



Photographing date: 0 0 1927 - 0 0 1933

Usage rights: Photographer

Place of origin: Jyväskylä

Additional location information

Reproduct. date

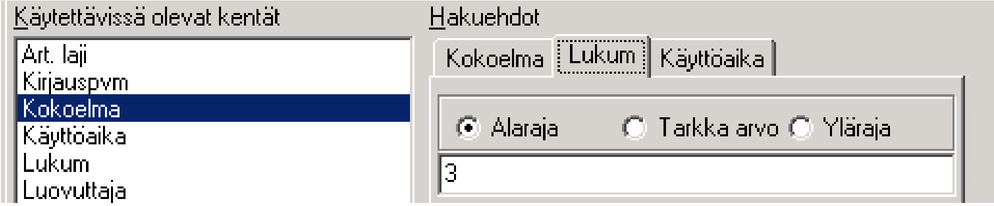
Reproduction producer

Additional reproduction information

- DateInterval is used to enter the (uncertain) photographing date. Unknown days and months are left empty (zeroes)
- Available punctuation marks for the interval can be browsed via drop down –list. Most commonly, standard interval mark "-" is used.
- New punctuation marks can be added only by staff with additional usage rights.

Querying Temporal Information

- For precise information (dates expressed in DATE datatype) exact match, or a given upper/lower bound can be used

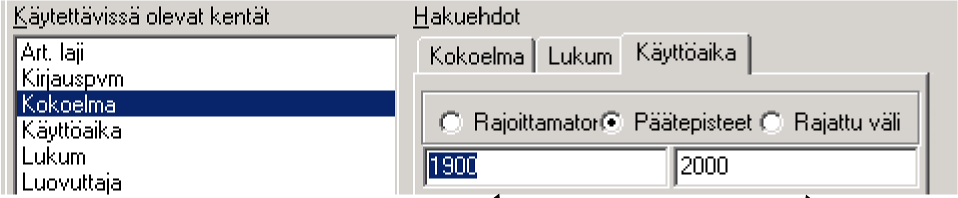


● Alaraja ○ Tarkka arvo ○ Yläraja

3

- For imprecise information, three search options are provided
  - **Contained interval (di)**: matches if both ends of the interval are contained within the query. The most restrictive search option.
  - **Overlapping interval (od)**: in addition to contained records, matches intervals that either overlap (or meet) with the query, or contain the query.
  - **Contemporary interval (ct)**: in addition to options above, includes intervals with potentially matching 0-years, such that it is possible that part of the result interval operlaps with the query.

Fields usable for search



○ Rajoittamaton ● Päätepisteet ○ Rajattu väli

1900 2000

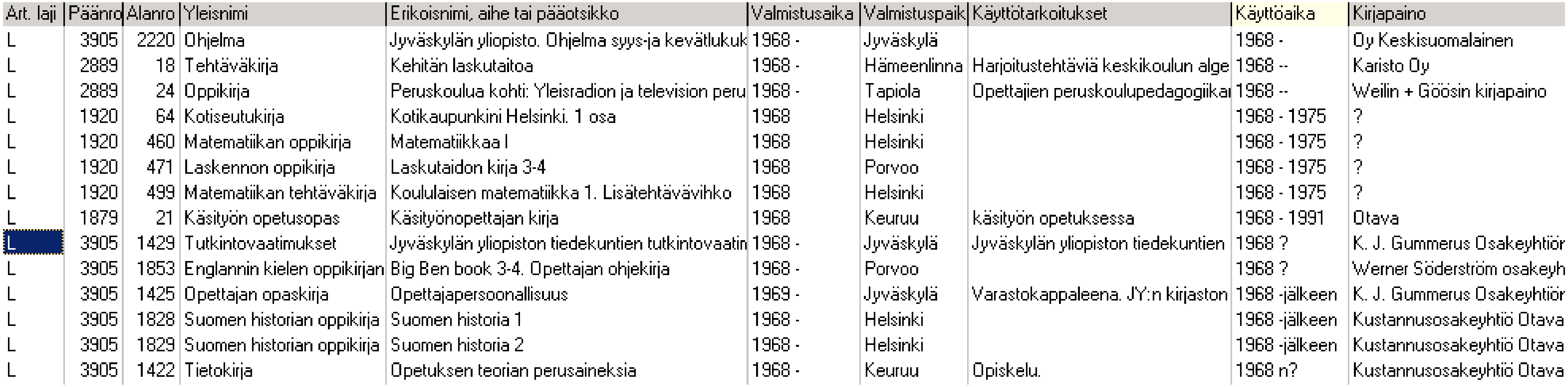
Matching options (di, od, ct)

Lower bound

Upper bound

Example Search Results

- Searching for books with usage date beginning at 1968…

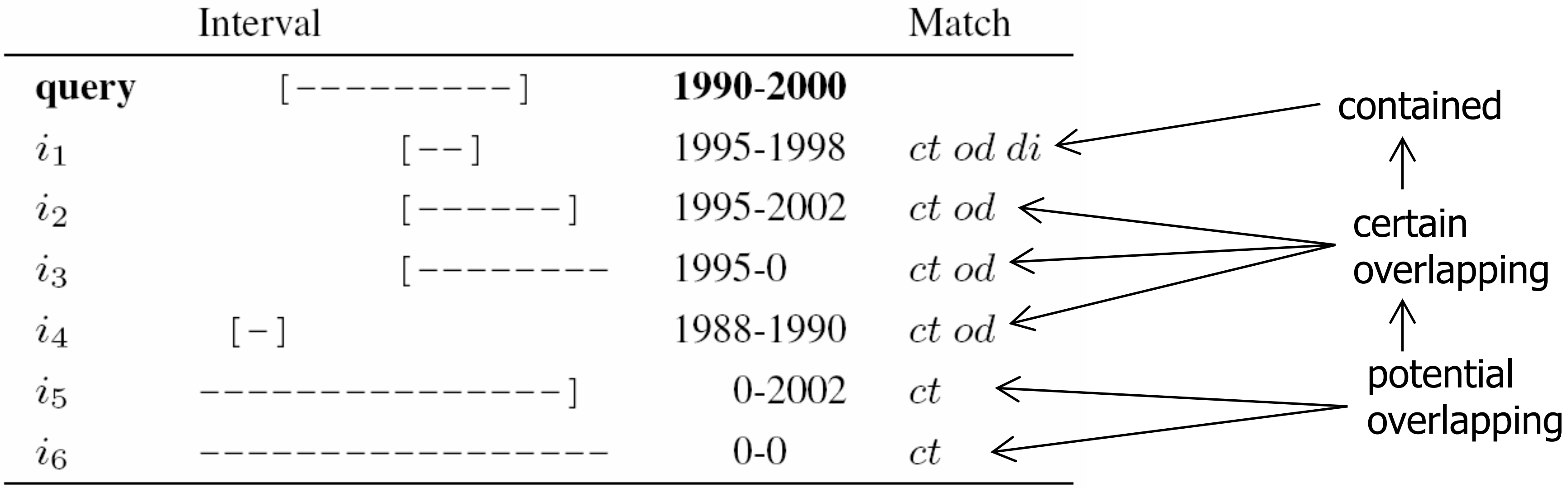


Art. laji	Päänro	Alanro	Yleisnimi	Erikoisnimi, aihe tai pääotsikko	Valmistusaika	Valmistuspaik	Käyttötarkoitukset	Käyttöaika	Kirjapaino
L	3905	2220	Ohjelma	Jyväskylän yliopisto. Ohjelma syys-ja kevätlukuk	1968 -	Jyväskylä		1968 -	Oy Keskisuomalainen
L	2889	18	Tehtäväkirja	Kehitän laskutaitoa	1968 -	Hämeenlinna	Harjoitustehtäviä keskikoulun alge	1968 --	Karisto Oy
L	2889	24	Oppikirja	Peruskoulua kohti: Yleisradion ja television peru	1968 -	Tapiola	Opettajien peruskoulupedagogiikai	1968 --	Weilin + Göösin kirjapaino
L	1920	64	Kotiseutukirja	Kotikaupunkini Helsinki. 1 osa	1968	Helsinki		1968 - 1975	?
L	1920	460	Matematiikan oppikirja	Matematiikkaa I	1968	Helsinki		1968 - 1975	?
L	1920	471	Laskennon oppikirja	Laskutaidon kirja 3-4	1968	Porvoo		1968 - 1975	?
L	1920	499	Matematiikan tehtäväkirja	Koululaisen matematiikka 1. Lisätehtävävihko	1968	Helsinki		1968 - 1975	?
L	1879	21	Käsityön opetusopas	Käsityönopettajan kirja	1968	Keuruu	käsityön opetuksessa	1968 - 1991	Otava
L	3905	1429	Tutkintovaatimukset	Jyväskylän yliopiston tiedekuntien tutkintovaatin	1968 -	Jyväskylä	Jyväskylän yliopiston tiedekuntien	1968 ?	K. J. Gummerus Osakeyhtiör
L	3905	1853	Englannin kielen oppikirjan	Big Ben book 3-4. Opettajan ohjekirja	1968 -	Porvoo		1968 ?	Werner Söderström osakeyh
L	3905	1425	Opettajan opaskirja	Opettajapersoonallisuus	1969 -	Jyväskylä	Varastokappaleena. JY:n kirjaston	1968 -jälkeen	K. J. Gummerus Osakeyhtiör
L	3905	1828	Suomen historian oppikirja	Suomen historia 1	1968 -	Helsinki		1968 -jälkeen	Kustannusosakeyhtiö Otava
L	3905	1829	Suomen historian oppikirja	Suomen historia 2	1968 -	Helsinki		1968 -jälkeen	Kustannusosakeyhtiö Otava
L	3905	1422	Tietokirja	Opetuksen teorian perusaineksia	1968 -	Keuruu	Opiskelu.	1968 n?	Kustannusosakeyhtiö Otava

- Start date, punctuation mark, and end date are compressed to one field.
- Punctuation marks are shown in the search results, but do not affect the search.

Detailed Search Example

- The query [1990-2000] is matched to different intervals
  - If contained search (*di*) is used only i_1 matches
  - For overlapping interval search, intervals i_2 to i_4 are matched as well, because they overlap or meet the query interval.
  - i_5 and i_6 are both matched with contemporary interval search, because with unknown values there is a possibility that they overlap.



	Interval		Match
query	[---------]	**1990-2000**	
i_1	[--]	1995-1998	*ct od di*
i_2	[------]	1995-2002	*ct od*
i_3	[--------	1995-0	*ct od*
i_4	[-]	1988-1990	*ct od*
i_5	---------------]	0-2002	*ct*
i_6	-----------------	0-0	*ct*

contained

certain overlapping

potential overlapping

- The terminology and abbreviations are adopted (but not equivalent) from Allen (during-inverse, overlapping/during) and Freksa (contemporary).

Problems with current approach

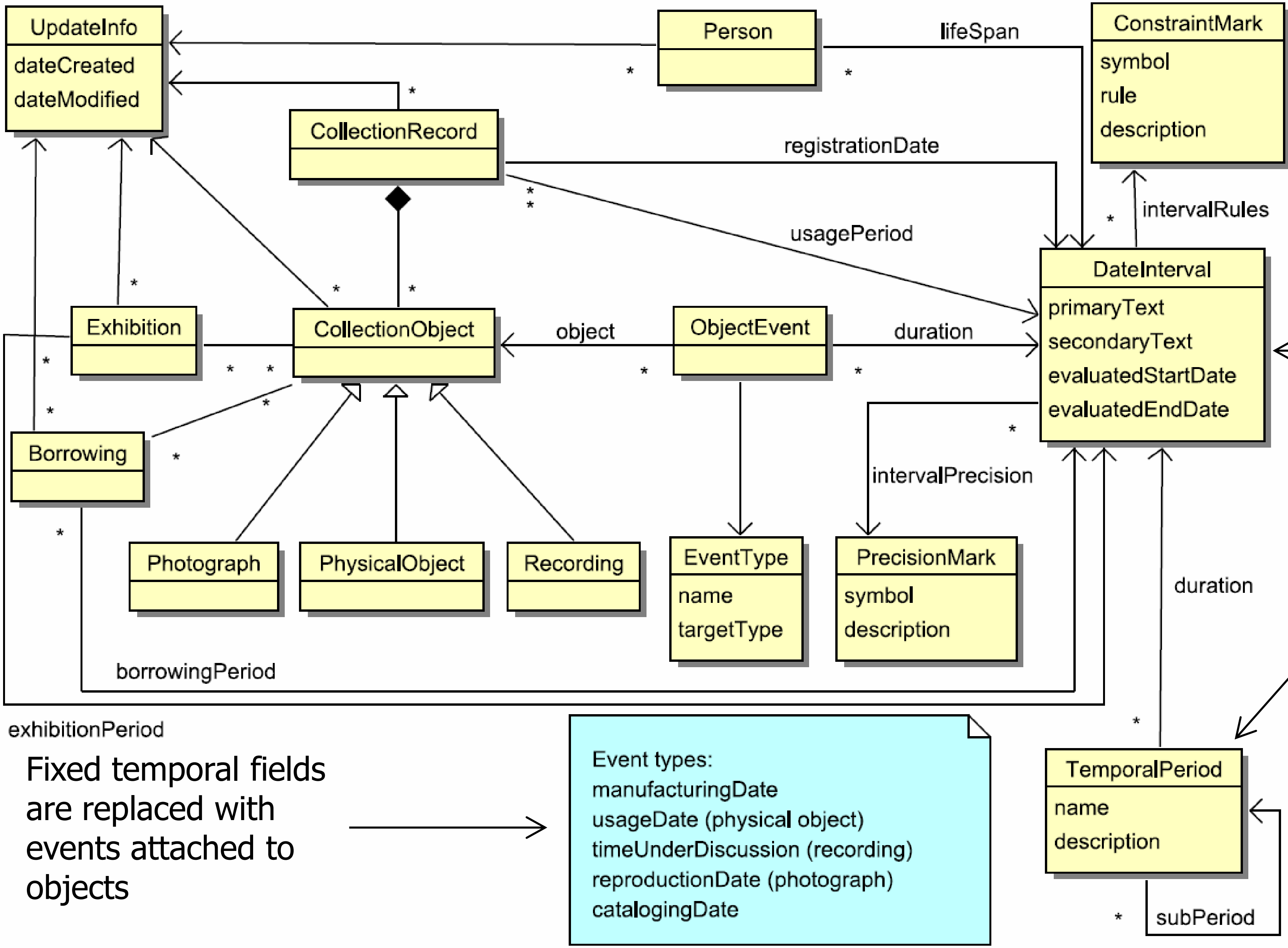
- Despite a few clean-up attempts, semantics in interval marks (and the words used) are not easily controlled
- Same query or data entry UI cannot be used for both precise (DATE) and imprecise (DateInterval) data fields
- No standard convention is enforced to present ”points” in time in DateInterval. A start date no interval mark can be interpreted as a point, this has not been used consistently.
- Definitions and user interface for different types of temporal queries is not intuitive to new end users
- Although the time representation in DateInterval is general-purpose, the database does not support a event-centric approach for object documentation (i.e. time information is ”hard-wired” to specific metadata fields, but cannot be used in an extensive way with user-defined roles.
- Some punctuation marks imply a combination of uncertainty and other, more “semantic” information.

```
- -- - - - ?? - ??-
- 1960 - d - ca.- -
c. / /april /spring
? -? -> 0 1937 -
1998 2001 2002 7
beg -beginning -beg
august before
february april -
after december
summer -d ? spring
beginning -d -d ca
? -d -d.beg d.end -
d? -end - decade -
decade- -decade?
decades -from
decade - from
decade? -end of
decade november ca-
until ca? circa -
circa circa-?
```

Requirements For a New Temporal Model

- Generalization of the temporal model such that both precise and imprecise information including both points and intervals are accounted for in same structure
- The data model should be brought closer to event-centric documentation to ease integration with new museum standards (i.e. CIDOC CRM)
- The number of available punctuation marks should be minimized to keep the model understandable and easy to apply
  - Symbolic information in punctuation marks should affect interval search
  - Punctuation marks related to uncertainty should be used with other symbolic information
- Could utilize ideas from time ontologies (e.g. query operators), but the representation should be physically in relational database.
  - Ease of integration to existing applications – minimize 3rd-party component usage to keep the application as self-contained and easy to install as possible
  - For maintainability, the implemented changes should affect the existing database and code as little as possible to enable smooth transition between versions.
- Introduction of named temporal periods for search terms

Proposed Schema for Temporal Entities



Event types:
manufacturingDate
usageDate (physical object)
timeUnderDiscussion (recording)
reproductionDate (photograph)
catalogingDate

PunctuationMark has been divided to two separate tables: PrecisionMark and ConstraintMark

Values entered by the user are separated from the DATE values that are used in query evaluation.

TemporalPeriod table can be used to specify named time periods in a hierarchical structure.

Fixed temporal fields are replaced with events attached to objects

Details For the New Model

- *DateInterval* table has been revised such that the values entered by the user are separated from the DATE values that are used in query evaluation.
- *PunctuationMark* table has been divided to two separate tables to represent symbolic constraints. *PrecisionMark* contains the symbols to present uncertainty (?, ca.), and *ConstraintMark* provides other symbols for custom granularities. The field *rule* states how the constraint mark affects the actual evaluated field values.
- *ObjectEvent* and *EventType* tables have been added to support event-centric documentation. Most DateInterval-based fields have been omitted from collection objects and consolidated to ObjectEvent table, allowing adding new event types without changes to schema. *TargetType* field in EventType provides guidelines to the user interface (e.g usage date is by default shown only with physical objects).
- Most of the DATE fields in the rest of the tables have been converted to DateInterval presentation to allow uniform search from the temporal fields.
- TemporalPeriod table can be used to specify named time periods (e.g. historical eras) in a hierarchical structure. Periods can be used for searching and data entry.l

Applying the New Model

- The actual evaluation of the user inputs (and possible application of the constraint rules) is left to the application since it would be very problematic to implement them using pure SQL in a portable way.
- The rule language for constraint marks is yet unspecified, but could be based on regular expressions and date arithmetic
- The constraint decade would pick all but last number from the beginning year entered by the user, and expand it such that the whole decade is covered.
  - The user input *1962 decade* would be evaluated to interval *[1960-01-01,1969-12-31]*.
- Also unknown days or months affect the evaluated dates: minimum and maximum dates in the given context
  - The user input *2011-02* would be evaluated to interval *[2011-02-01,2011-02-28]*.
- Alternate dates (OR relationship) are not modeled as a constraint mark, but by modifying the schema such that an indeterminate number of named object events can be attached to the collection object and treated as alternates.
  - This is in line with conventions used in CIDOC CRM and semantic web applications in general.

Evaluation

- Since the model is not yet actually implemented and not tested with end users, the evaluation is at best preliminary.
- The database schema is still highly normalized, but because of the additional joins with events, the query performance will be somewhat slower than before
- The decision to store the intervals in a separate table with idiosyncratic rules and external code can be regarded as a disadvantage from the portability standpoint. However, the old model was even less portable regarding uncertain intervals and did not use DATE datatype at all.
- The most critical limitation of the data model is the lack of semantics within the precision marks, but there are multiple practical issues if they were used in the query evaluation:
  - Ranked searches can not be naturally implemented within a relational database
  - It is not clear what would be the appropriate ”fuzziness” and shape of the fuzzy set used to describe the interval
  - The specification of the intervals would be laborious to the end user.

Conformance to Museum Standards

- The new data model implements most of the requirements specified in the common library and museum standards regarding the presentation of temporal information, although some of the mechanisms are used or named differently.
  - Most of the temporal information specified with *SPECTRUM* standard can be represented with the new model. For simple expressions (Late 19th century or early 20th century), constraint marks can be used to similar effect with the added benefit of making the expressions searchable. However, the notion of qualifiers to explicitly mark up the probable deviation for start and end instants (e.g. uncertain start date, *±10 years*) is not supported in Duo.
  - *CIDOC CRM* contains a conceptual hierarchy related to different kinds of events (e.g. birth, creation, transformation) that can partially be presented with EventType table in Duo. CRM class E52 Time-Span is analogical to DateInterval in Duo, with the addition of qualifiers, and allowing additional temporal relations to be defined between the intervals and other classes. The temporal model defined in CIDOC is clearly more expressive compared to the model in Duo, but very involved to implement or mark up the data.
  - The conventions for uncertain dates as defined in *CCA's Rules for Archival Description* are semantically very close to the model defined in Duo with differences in notation. The authors consider it a good compromise between expressivity and easy markup. For example, *probable 17th century* can be marked as "17-?". In Duo, similar effect can be accomplished with constraint marks.

Comparison With Other MIS

- Musketti is a popular collection management system used by many museums in Finland. According to the documentation, its the temporal model has many issues compared to the one in Duo, such as *four* separate mechanisms for marking up intervals and lacking normalization. Some of the temporal fields are stored as strings and thus, lack numeric searchability without manual markup of numeric dates.
- Polydoc is another commercial collection management system used in Scandinavia. Polydoc contains a simple interval-based mechanism for marking up temporal data, along with optional textual information in a separate field. While easy to use, it is clearly less expressive compared to the Duo data model.
- Emerging open source alternatives (CollectiveAccess, CollectionSpace) need additional consideration since they claim to support user-defined schemas and museum standards. However, we were not able to review the systems in detail in this paper.

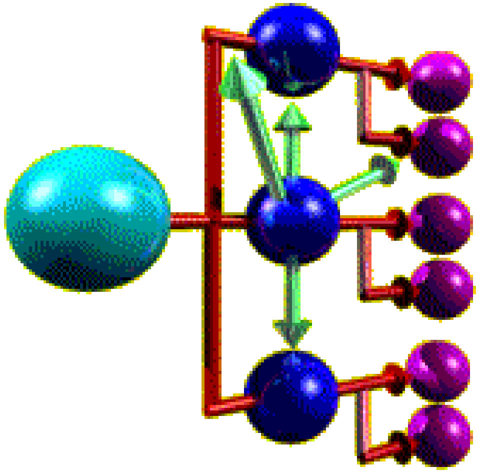
Addressing Reviewers’ Notes

- The text was slightly shortened and focused, especially concerning the literature review. More effort was put to evaluation.
- Nonstandard search constraints terminology (*ct, od, di*) was elaborated in more detail.
- The term *"object lifecycle"* caused some confusion with a reviewer. We have replaced it constantly with *event-centric documentation*.
- While we are aware that significant amount of research concerning both database technology and temporal information has been conducted during last decades and attempted to review some of it, we were confused with a reviewer pointing to “significant European authors considering this topic”. We welcome additional directions about the relevant papers we have missed.

Conclusion & Further Research

- Culture historical information provides a rich and challenging domain for data management, both from temporal and general perspective.
- A collection management system used in JYU Museum was introduced and problems with representing and retrieving temporal information were identified.
- A new temporal model accounting different representations, uncertainty, and event-centric documentation was roughly sketched.
- Future research involves the implementation of the model in a relational database and transformation of the old data from the production database.
- The model and rule language must be specified in more detail along with potential user interface in cooperation with end users
- Treating the MIS as an temporal database is an attractive prospect. For example, a timestamp presenting the latest change in a given record is alrealy used since this allows querying for latest changes in the database. Enhancing this to explicit audit trail would allow undo operations and documents data cleaning.
- Mapping rules and temporal periods for culture-sensitive (uncertain) temporal information and generalizing the model to different calendars would also be an interesting prospect.

Thank You!



Further information:

•Anneli Heimbürger, Dr. Tech
http://users.jyu.fi/~anheimbu/
anneli.a.heimburger@jyu.fi

•Miika Nurminen, M.Sc
http://users.jyu.fi/~minurmin/
minurmin@jyu.fi

Source: ISO/IEC JTC1/SC18/WG8 N1920
Information technology — Hypermedia/Time-based Structuring Language (HyTime)
http://www1.y12.doe.gov/capabilities/sgml/wg8/document/n1920/html/n1920.html