



NATIONAL INSTITUTES OF HEALTH
enterpriseARCHITECTURE

EATS – August 3, 2006

Enterprise Standards and Data
Representation



NIH Enterprise Information Technology Architecture
Contact: enterprisearchitecture@mail.nih.gov



Objectives

- Provide an overview of Enterprise Data Standards and the categories of standards that may be created within NIH.
- Provide an overview of a proposed approach to the representation of data standards as used in NIH Request for Comment (NRFC) 0002 – Person Name.
- Discuss the requirements for effective representation of data standards and define next steps for refining NIH's approach.





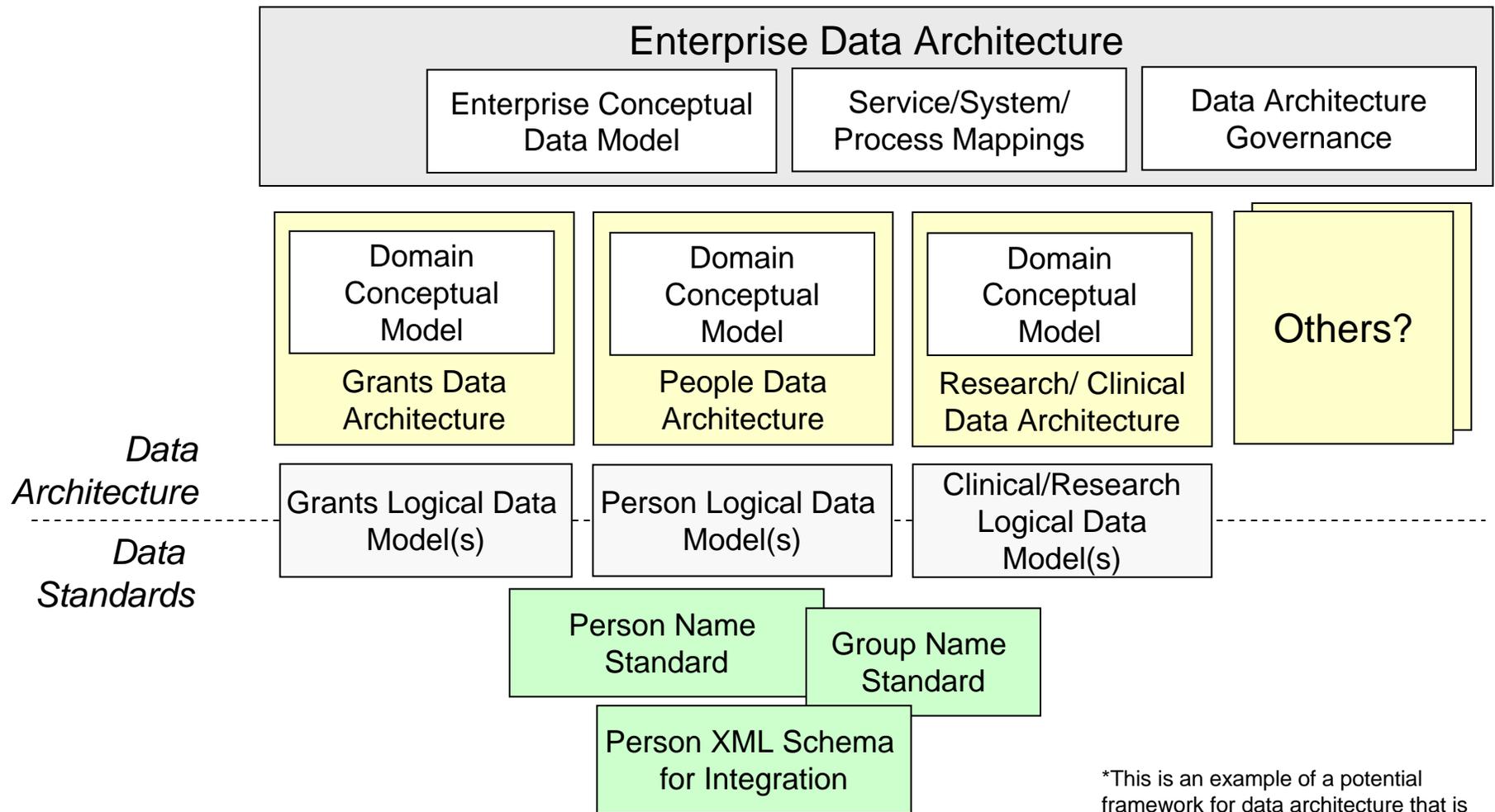
Data Representation Challenges

- All Data Standards are not created equal
 - Different types of data standards may demand different representations
- A variety of audiences need to be served by data standards
 - Architecture and designers
 - Developers
 - Business SMEs?
 - End users?
 - Business Owners?
- Data Standards may need to be represented differently for different audiences
- Data Standards and Enterprise Data Architecture are very much interrelated, but are not the same thing.





Overview – Notional Data Architecture Framework*



*This is an example of a potential framework for data architecture that is currently under development and included here for discussion only.





Differentiating between Data Architecture and Data Standards - *Categories of Artifacts*

Enterprise Data Architecture Models and Plans

- Define direction for the NIH enterprise.
- Specify current state, target state, and transition.
- Support allocation of governance responsibility for data elements.
- Allow for the mapping of data architecture elements to other layers within the architecture
 - Business Architecture
 - Technology Architecture
 - Application Architecture
- Support impact analysis.

Data Standards

- Provide detailed guidance for implementers with respect to the structure and use of information within NIH.
- Bridge between architecture and solution design.
- Provide detailed technical specifications for formats used in data management and data exchange.





Types of Data Standards - Illustrative

Logical Data Standards

- Define the overall structure and information content of key NIH data entities.
- Define the business rules for describing data.
- Provide logical models that will support further development of physical models in support of a specific solution.

Implementation Data Standards

- Define implementation-specific standards. For example:
 - Definition of XML formats for data exchange.
 - Definition of standards for data captured within AD.
 - Mapping of person data to specific LDAP repository structures.
- May be accompanied by technical artifacts such as XML schemas or physical data models.





Data Architecture Artifact Examples by Category

Category	Artifact	Description	Validation/Approval (Notional)
Data Architecture	Enterprise Conceptual Data Model and Entity List	High level listing of key NIH business enterprise data objects and their relationships.	“Domain Team” (Guided NRFC Process)/ARB Approval
Data Architecture	Subject Area Model	Groupings of Enterprise CDM objects to reflect relatedness of information and alignment with business ownership.	“Domain Team”/ARB Approval
Data Architecture	Data/Application (Service) Mapping	A mapping of key data enterprise data objects to the applications and services that use them including CRUD.	“Domain Team”/ARB Approval
Data Architecture	Data/Business Process Mapping	A mapping of data objects to the business processes in which they are used.	“Domain Team”/ARB Approval
Data Architecture	Data Ownership Matrix	Assignment of ownership of data objects to appropriate business owners.	Business Owners/ITWG Approval
Data Architecture	Data Governance Strategy	Overall strategy for data governance within NIH including the responsibilities of data owners and stewards and the allocation of decision authority across roles.	ITWG Approval





Data Architecture Artifact Examples by Category (cont'd)

Category	Artifact	Description	Validation/Approval (Notional)
Logical Data Standard	Person Conceptual/Logical Data Model	More detailed model of the person domain intended to codify the business rules and structure for information related to people and organizations within NIH.	Working Group/ NRFC/ARB Approval
Logical Data Standard	Grants Conceptual/Logical Data Model	More detailed model of the grants domain intended to codify the business rules and structure for information related to grants within NIH.	Working Group/ NRFC/ARB Approval
Technical Data Standard	XML Schemas for Application Integration	An XML schema defined in XSD or another schema specification language intended for use in system-to-system integrations leveraging either message-oriented middleware, web services, and other integration approaches.	NRFC/ARB Approval
Technical Data Standard	LDAP Repository Schemas	A schema for an LDAP repository defined formally as specified in IETF RFC 2252 (http://www.ietf.org/rfc/rfc2252.txt) and accompanying documentation	NRFC/ARB Approval
Technical Data Standard	Field Format Standards (e.g., AD Group Names, Person Name)	A formal definition of the standards for a particular field or key data element that is used across NIH. An example of such a standards is NRFC0005 – “Unique Identifier for People, the NIH ID Number”	NRFC/ARB Approval





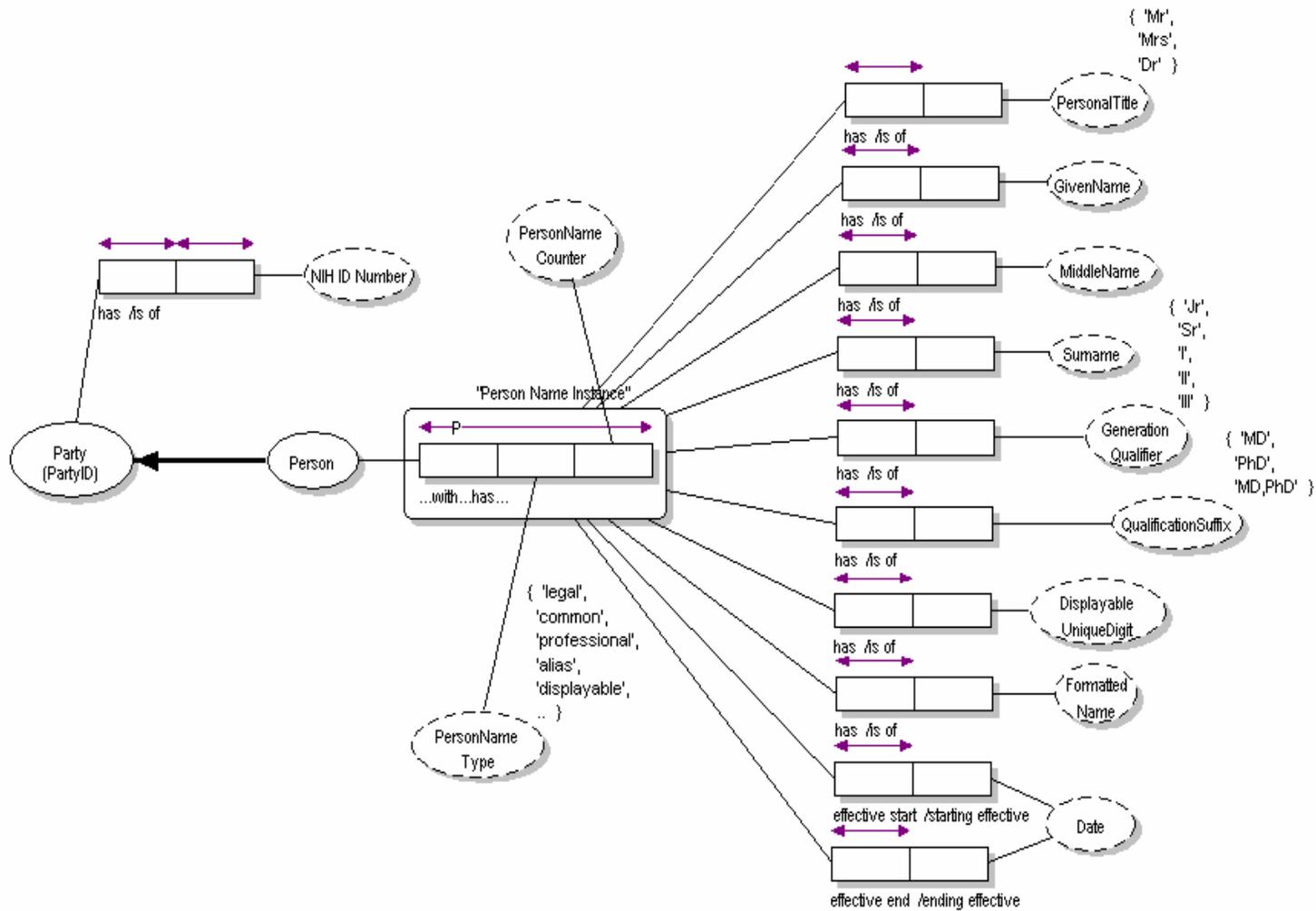
NRFC 0002 – Person Name Approach Overview

- The Person Name NRFC was developed using Natural Language Modeling (NLM) techniques and Object Role Modeling (ORM).
 - NLM and ORM provide a means to specify the business rules of managing information through sentences and in visual models.
 - ORM models provide a more rigorous means of specifying relationships than in ERDs.
- NRFC 0002 is in draft format and is undergoing another revision prior to submission to the ARB for approval.



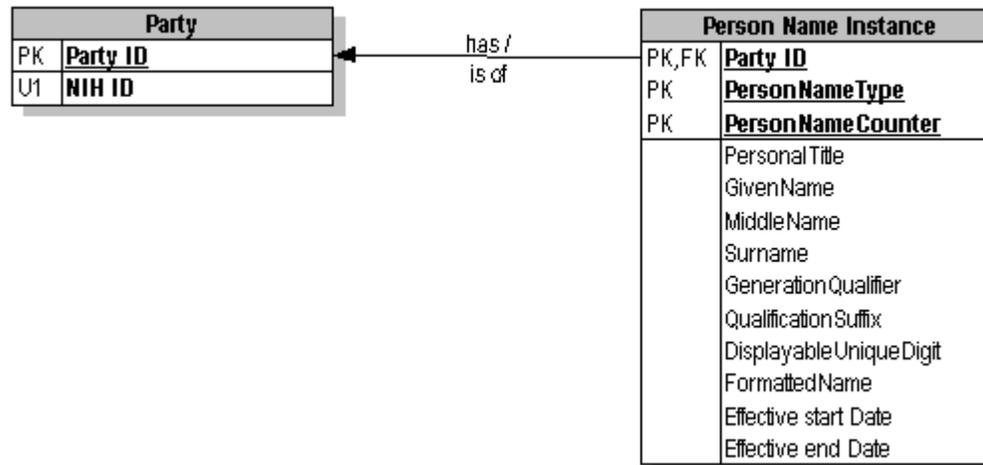


Representing Data Standards – Data Models (ORM)

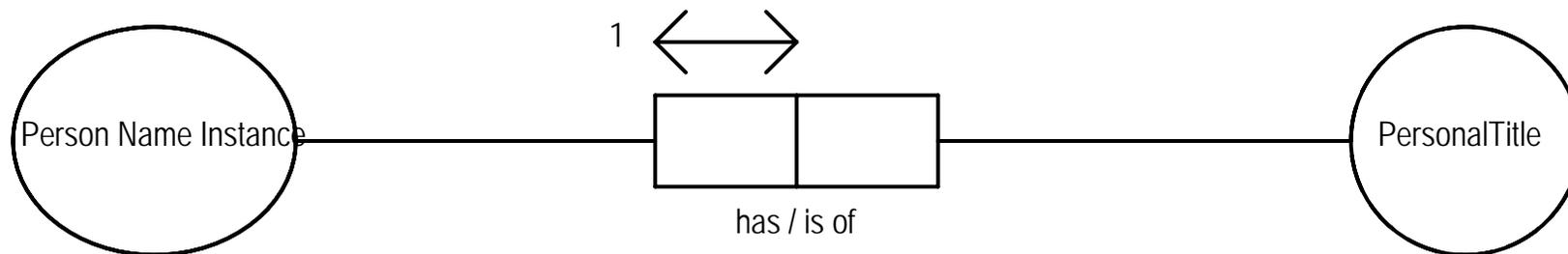




Representing Data Standards – Data Models (ERD)



Representing Data Standards – Describing Attributes



Record: 16
Fact: Person Name Instance has PersonalTitle
Inverse fact: PersonalTitle is of Person Name Instance
Mapping Option: Does not result in a composite type
Constraints: 1
1: Each Person Name Instance has at most one PersonalTitle.
External rules: None
Examples: None
Defined on: Page 1 (NRF0002 Data Model 1)



Representing Data Standards – Expressing Business Rules

Person Name Type Derivation Rules

Name Type	Max Current #	Mandatory Name	Derivation Rules for Name Types
Legal	1	Y	Name concatenated from allowed populations
Common	1	Y	Name concatenated from allowed populations - Legal Name is returned if none exist [Legal Surname mandatory]
Professional	1	N	Name concatenated from allowed populations - Legal Name is returned if none exist
Alias	∞	N	Name concatenated from allowed populations
Badge Name	1	N	Concatenated: Given Legal Name + Middle Legal Initial + Legal Surname except when Given Legal Name is an Initial, then Given Legal Name + Middle Legal Name + Legal Surname
System Account Name	1	N	Legal Surname + first initial of Common Given Name or if none Legal Common Name + if needed first initial Common Middle Name or if none Legal Middle Name + if needed Digit (Digit stored and reused in generating displayable name, active main email address and other Email address (if needed))
Displayable Name	1	N	Legal Surname + ", " + Common Given Name or if none Legal Given Name + " " + Common Middle Name or if none Legal Middle Name + " " + (possibly generation qualifier?) + Digit + Organization + Person Status
Main Email Address	1	N	Common Given Name + "." + Legal Surname + if needed a digit + "@nih.hhs.gov"
Other Email Address	∞	N	User Defined (may have limitations)





Representing Data Standards – Expressing Business Rules

Name Type	Allowed Population									
	Personal Title	Given Name	Middle Name	Surname	Generation Qualifier	Qualification Suffix	Displayable Unique Digit	Formatted Name	Effective Starting Date	Effective Ending Date
Legal	-	Y	Y	Y	Y	-	-	Y	Y	Y
Common	-	Y	Y	Y	Y	-	-	Y	Y	Y
Professional	-	Y	Y	Y	Y	-	-	Y	Y	Y
Alias	Y	Y	Y	Y	Y	Y	-	Y	Y	Y
Badge Name	-	-	-	-	-	-	-	Y	Y	Y
System Account Name	-	-	-	-	-	-	Y	Y	Y	Y
Displayable Name	-	-	-	-	-	-	-	Y	Y	Y
Main Email Address	-	-	-	-	-	-	-	Y	Y	Y
Other Email Address	-	-	-	-	-	-	-	Y	Y	Y





Unique Digit Rule

The system account name, displayable name and the main email address for a person must be unique. If the three derived names are not unique then a displayable unique digit must be determined. The displayable unique digit must be the same for all three names.

Rule for determining the displayable unique digit: Determine the smallest digit (if needed) to make each name unique. Of these digits, test the largest one on the other two names. If all names are unique then the displayable unique digit has been found. If a name is not unique then sequentially increase the digit until all names are unique. The final digit will be the displayable unique digit. This digit will be added to all three names (even if the original name is unique). [The Main Email Address must be unique across all email addresses.]





Discussion

- What are the audiences for data standards developed as part of the NIH EA?
- What additional contextual information would be useful as part of data standards? E.g.,
 - Intended Audience
 - Recommended Use
 - Limitations
 - ...
- What notations in addition to ORM and ERD would be useful?
- How should data standards be categorized?
- How should data standards be validated?
- Other questions?





Next Steps

- Aggregate feedback from discussion and NRFC development.
- Develop guidance for Data Standard development.
- Others?

