

DEPARTMENT OF
INFORMATION
ENGINEERING
UNIVERSITY OF PADOVA



A Set-Based Approach to Deal with Hierarchical Structures

Gianmaria Silvello

PhD School in Information Engineering

Information and Communication Science and Technology

Series: XXIII

Supervisors: Prof. Maristella Agosti and Dr. Nicola Ferro

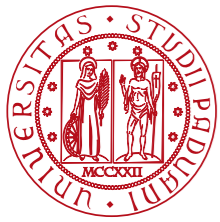
Information Management Systems (IMS) Research Group,
Department of Information Engineering,
University of Padua, Italy

PhD Defense Presentation
19 April 2011, Padua, Italy



Outline

- Background
- Research Question and Context
- The NESTOR Model
- The NESTOR Prototype
- Final Remarks and Future Work



Background



Hierarchies

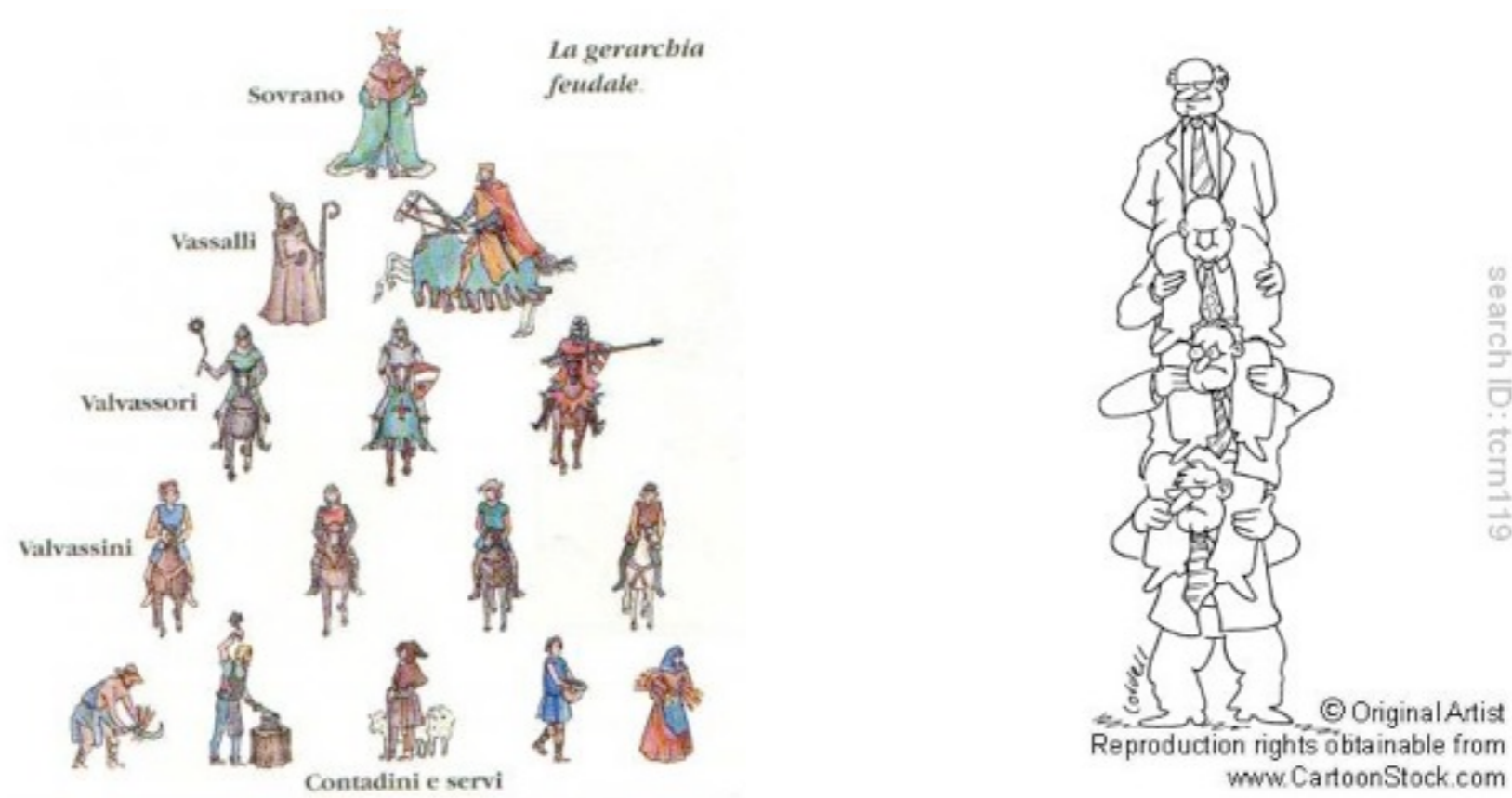
- Hierarchies are part of our common experience in the physical as well as in the living and social worlds [\[CourceauEtA106\]](#).
- Hierarchy is a deep and powerful concept that has a polysemous semantic [\[Simon62\]](#).

- Hierarchies are part of our common experience in the physical as well as in the living and social worlds [CourgeauEtA106].
- Hierarchy is a deep and powerful concept that has a polysemous semantic [Simon62].



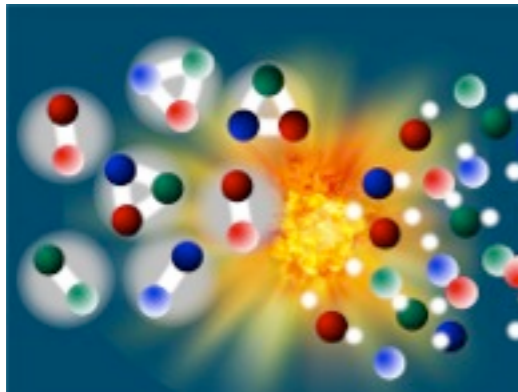
Inclusion hierarchy: recursive organization of entities: the “chinese box” metaphor.

- Hierarchies are part of our common experience in the physical as well as in the living and social worlds [CourceauEtA106].
- Hierarchy is a deep and powerful concept that has a polysemous semantic [Simon62].

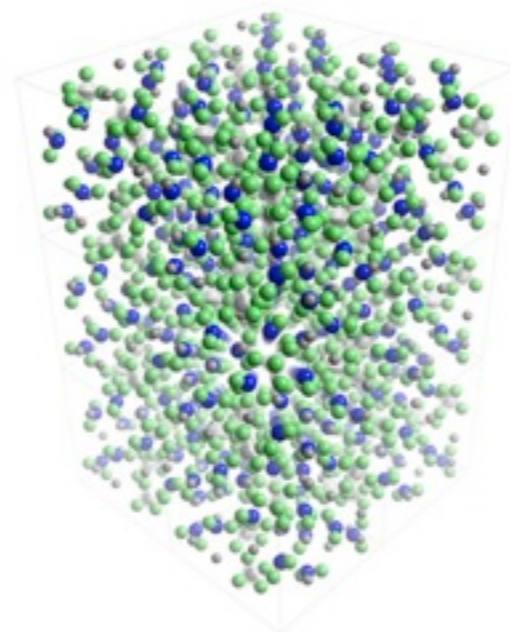


Control hierarchy: social organization - who gives orders to whom; a control system in which every entity has an assigned rank.

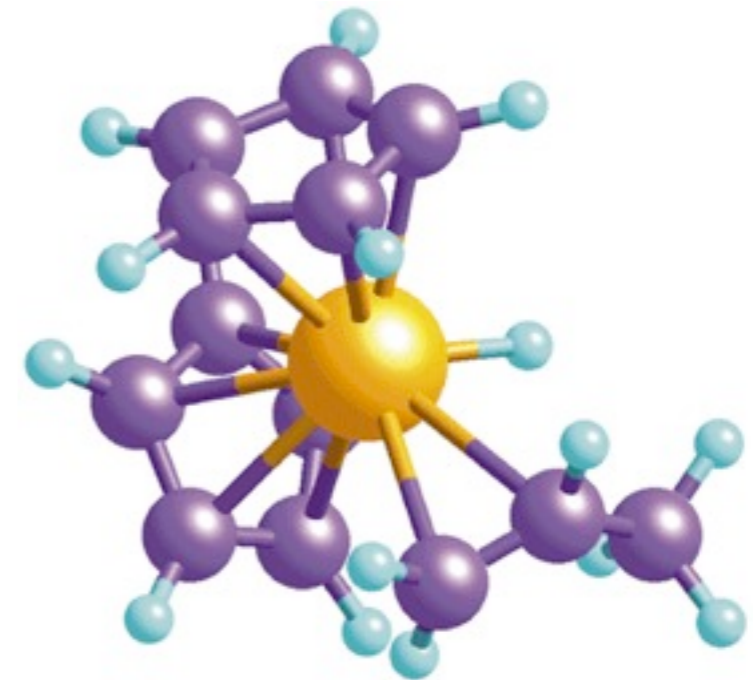
- Hierarchies are part of our common experience in the physical as well as in the living and social worlds [CourceauEtAl06].
- Hierarchy is a deep and powerful concept that has a polysemous semantic [Simon62].



Elementary Particles

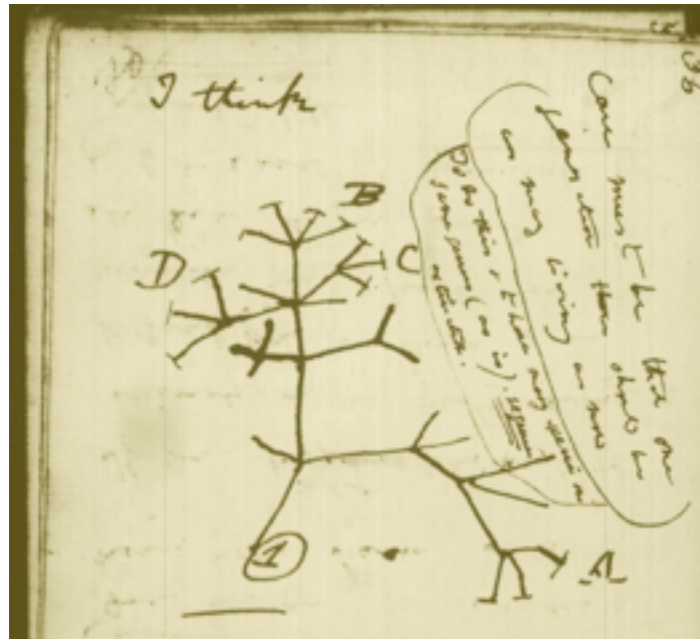


Atoms



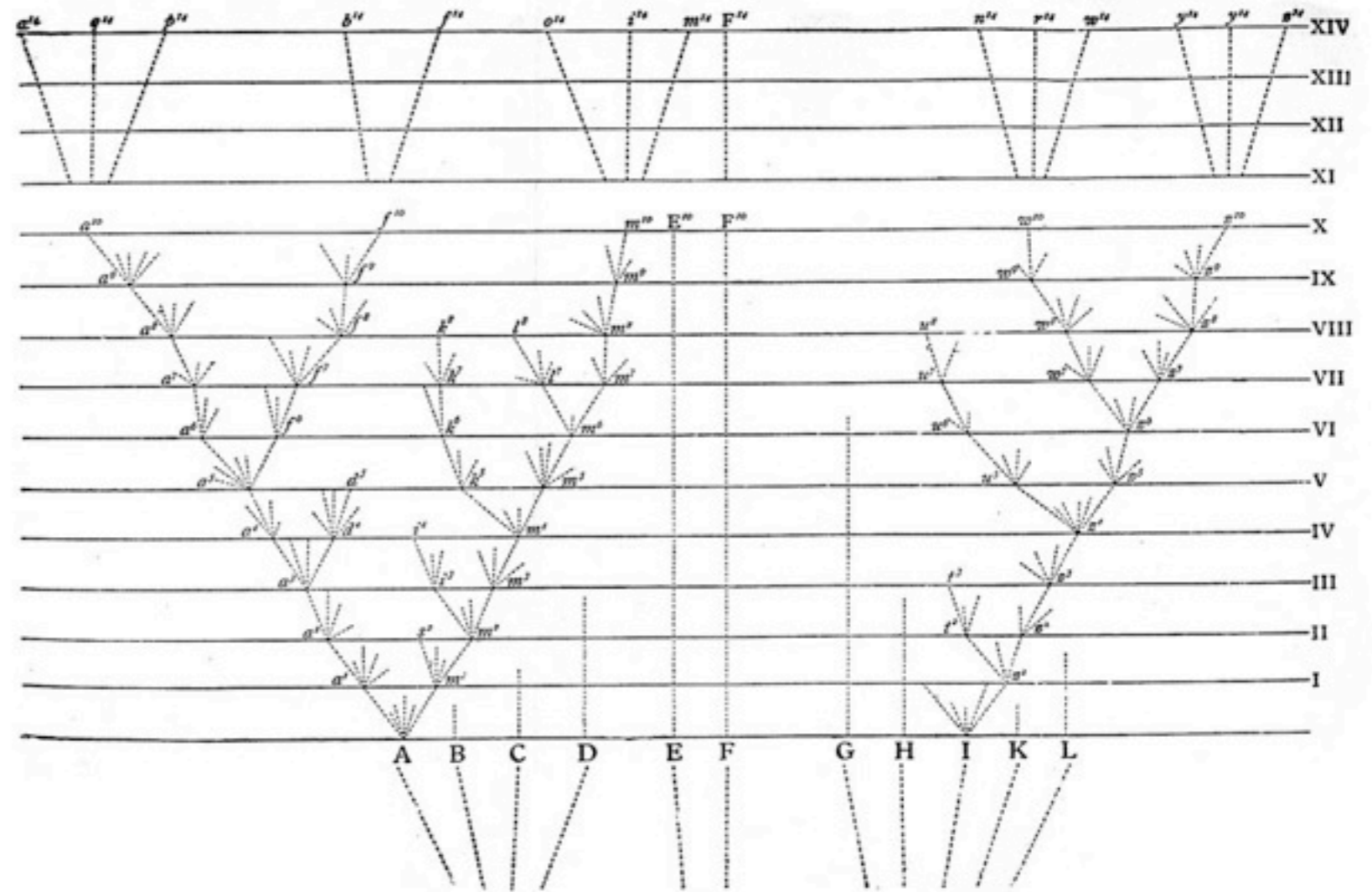
Molecules

Level hierarchy: Each level is characterized by a particular spatiotemporal scale for its associated entities and for the processes through which the entities at this level interact with one another.

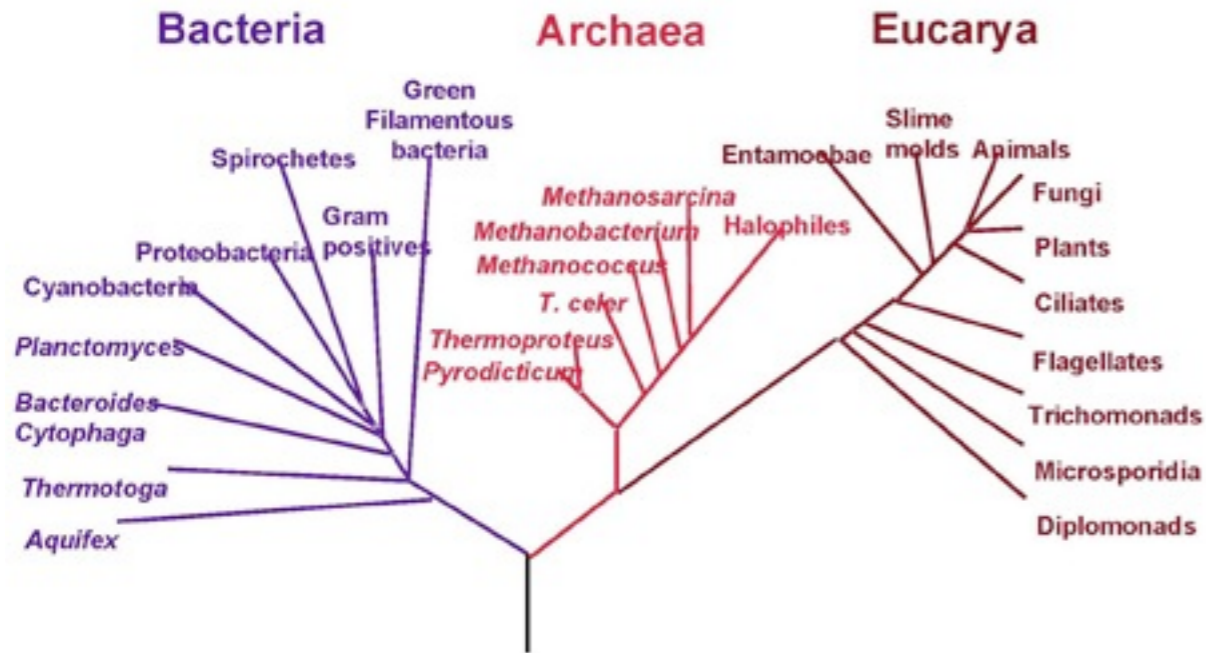


Darwin's Evolutionary Tree [Darwin1859].

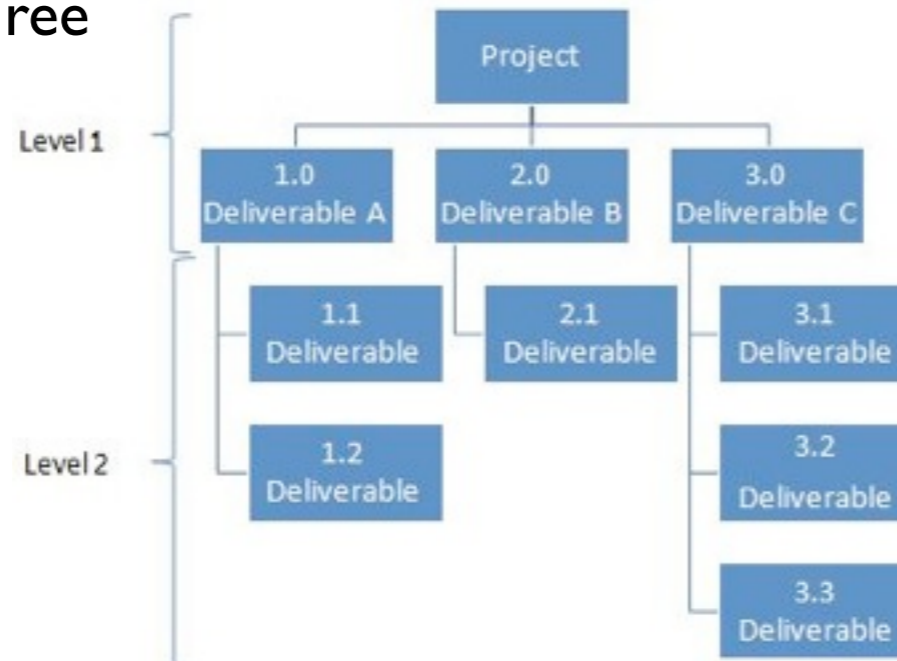
“It is an odd looking affair, but *indispensable* to show the nature of the very complex affinities of past and present animals” [Moretti05].



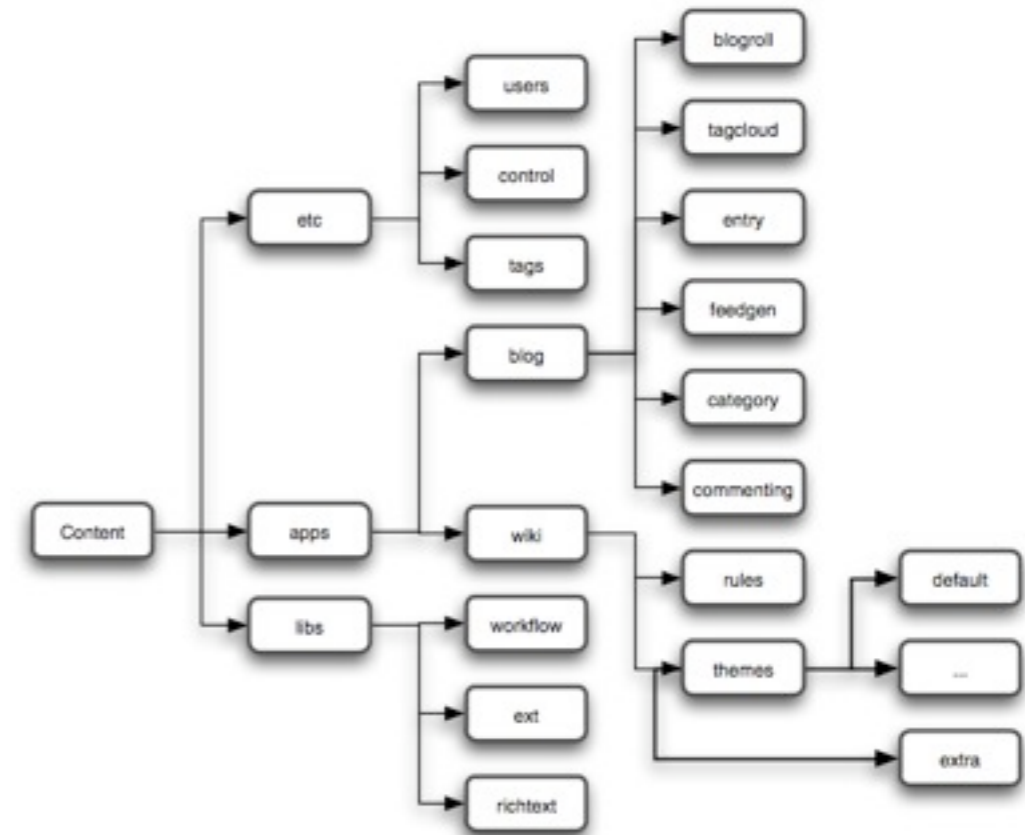
Phylogenetic Tree of Life

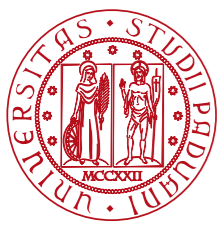


Organization Tree



File System Tree





The Tree Data Structure

- The tree representation is formally defined as a proper data structure and it is one of the most important data structure in computer science [Knuth97].

Formal Definition in Graph Theory [Christofides75]:

Let $V = \{v_1, \dots, v_n\}$ be a set of nodes and the set E is a mapping of the set V in V , $E : V \rightarrow V$, thus $T(V, E) = T(V, V \times V)$; E is defined as a set of couples $\{v_i, v_j\}$ where $v_i, v_j \in V$ such that v_i is connected to v_j and thus v_i is the parent of v_j . If $T(V, E)$ is:

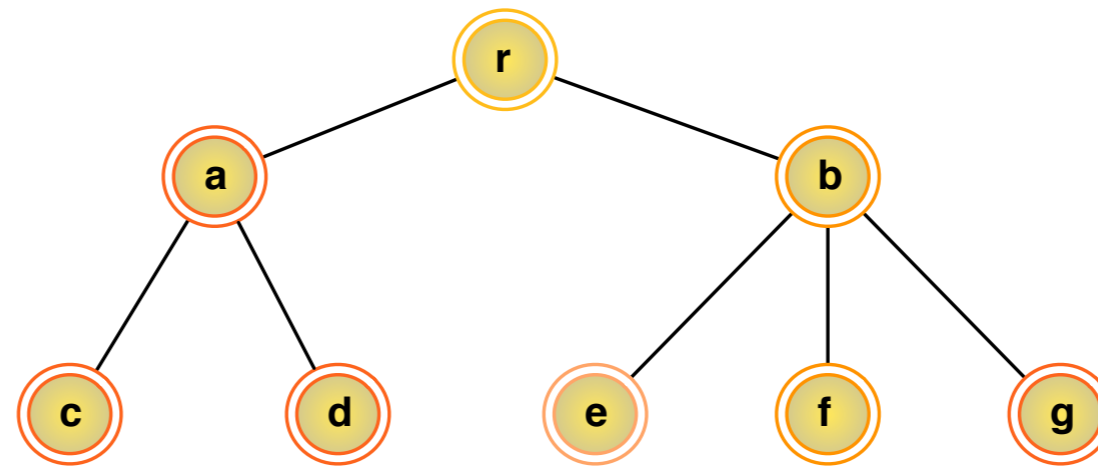
- (i) A connected graph of n vertices and $(n - 1)$ links,
- or (ii) A connected graph without a circuit,
- or (iii) A graph in which every pair of vertices is connected with one and only one elementary path,

Then $T(V, E)$ is a tree.

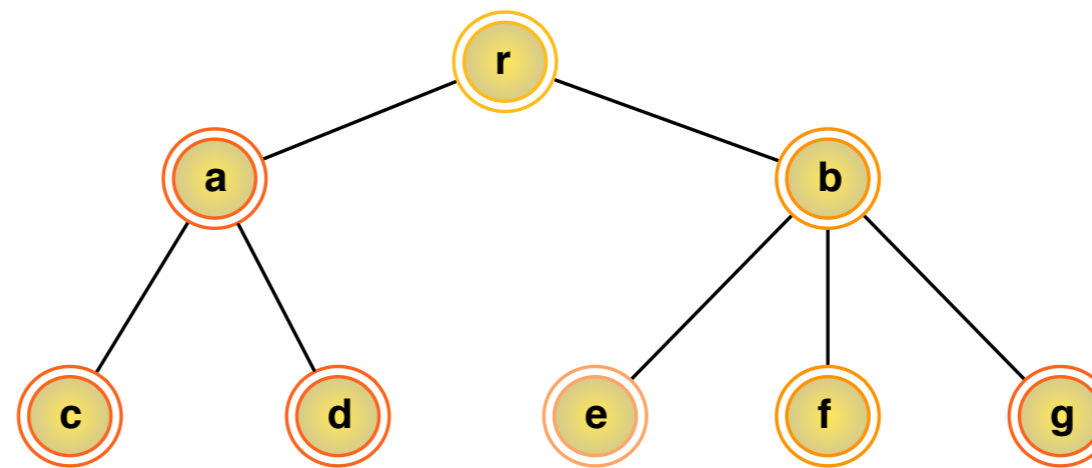


Research Question and Context

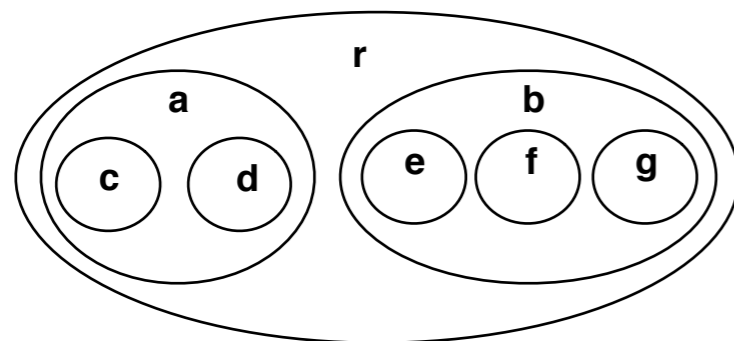
- When we think of a hierarchy in computer science we think at it as represented by a set of nodes and edges.



- When we think of a hierarchy in computer science we think at it as represented by a set of nodes and edges.



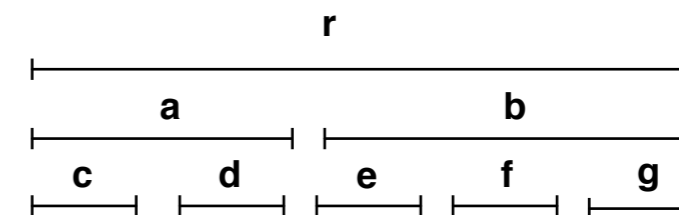
But it is also possible to point out some alternative representations [\[Knuth97\]](#):



Nested Sets

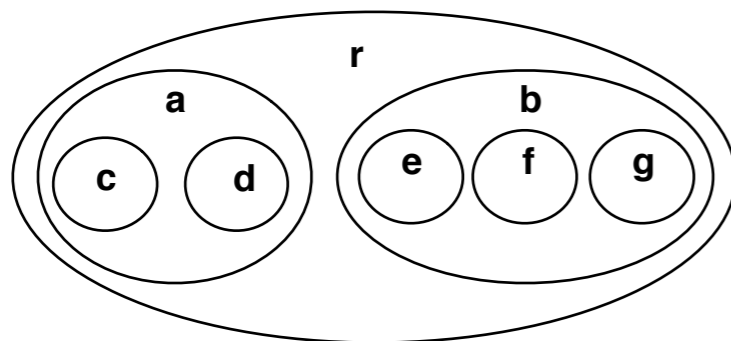
$r(a((c)(d)) b((e)(f)(g)))$

Nested Parenthesis



Nested Intervals

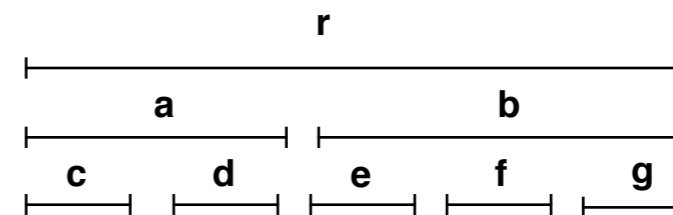
But it is also possible to point out some alternative representations [\[Knuth97\]](#):



Nested Sets

$r(a((c)(d)) b((e)(f)(g)))$

Nested Parenthesis



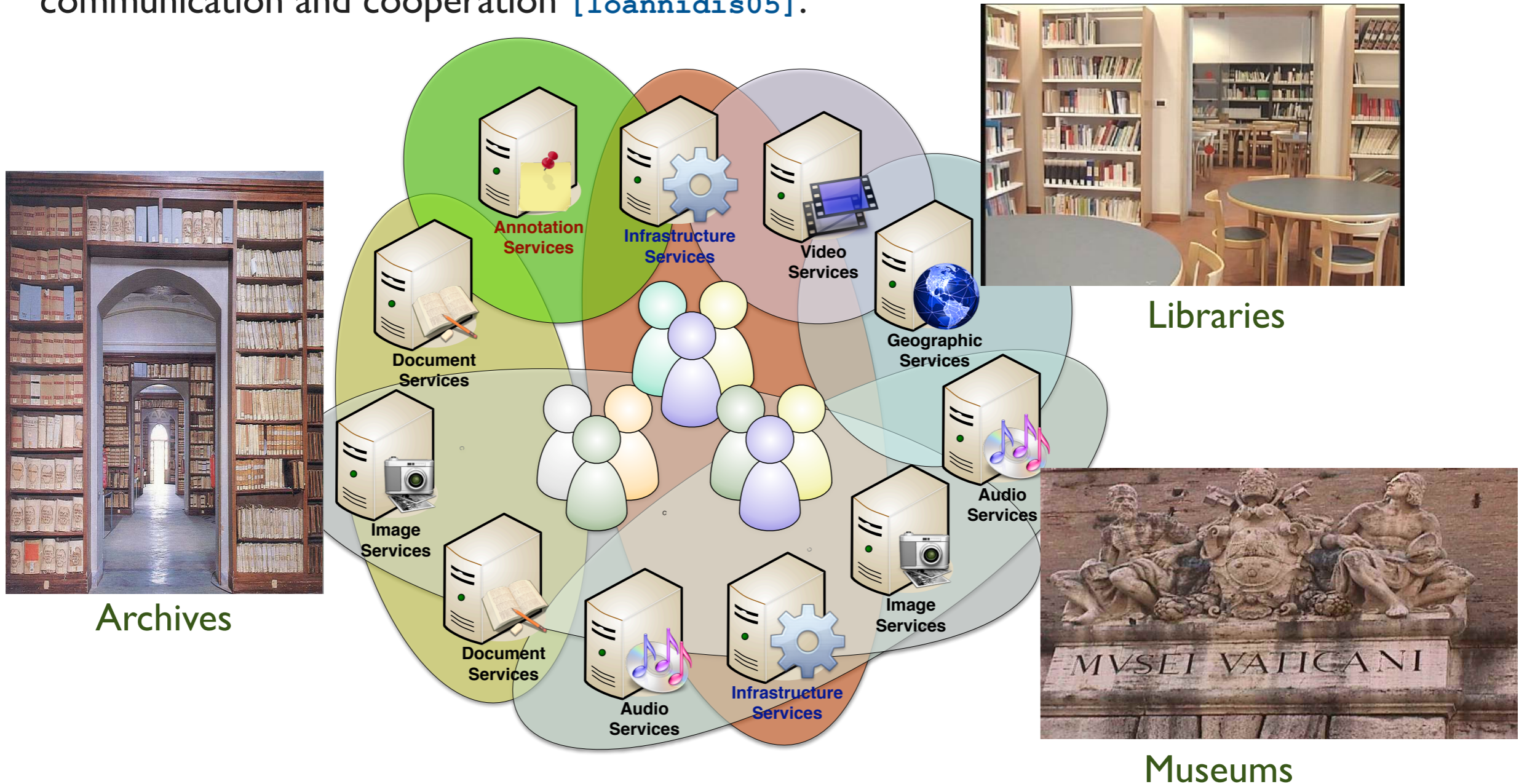
Nested Intervals

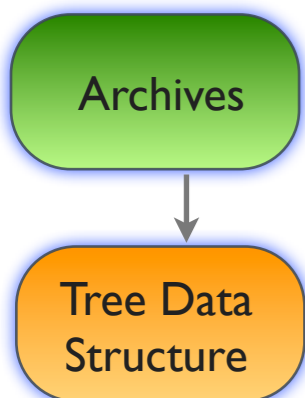


Are they just representations?

Or could we formally define new data structures based on these intuitive ideas pointing out new properties?

- A **Digital Library** is a **collection of information** that is both **digitized** and **organized**. A digital library can be **searched** for any phrase, it can be **accessed** all over the world [Lesk97]. Digital Libraries are user-centric systems devoted to communication and cooperation [Ioannidis05].

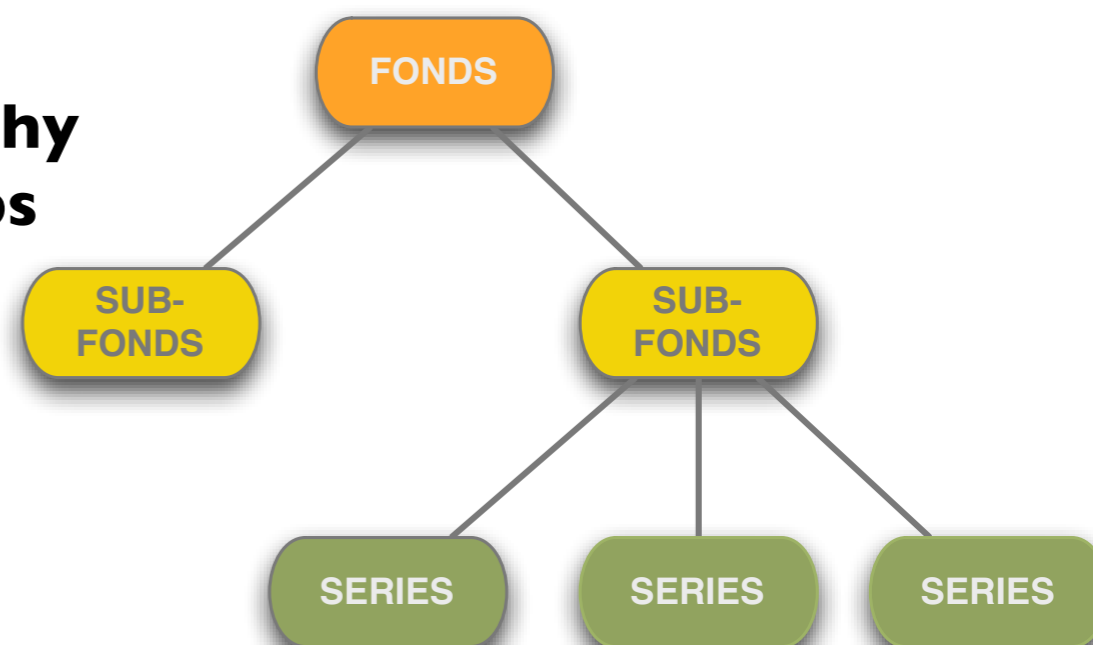




- An archive represents the **trace of the activities** of a physical or juridical person in the course of their business which is preserved because of their continued value [\[Duranti98\]](#).
- Archives keep the **context** in which their records have been created and the **relationships** among them in order to preserve their informative content and provide understandable information over time [\[Duranti98\]](#).



Archival descriptions constitute a **hierarchy which represents the relationships** among more general and more specific archival units [\[Pearce-Moses05\]](#).





The NESTOR Model

State of the art

**Graphical
Representation**
[Knuth97]



**Integer
Encoding**
[Celko00]

Nested Intervals
[Tropasko05]

The NESTOR Model

**The Nested Set
Model
NS-M**

**The Inverse
Nested Set Model
INS-M**

**Properties of the Set Data
Models**

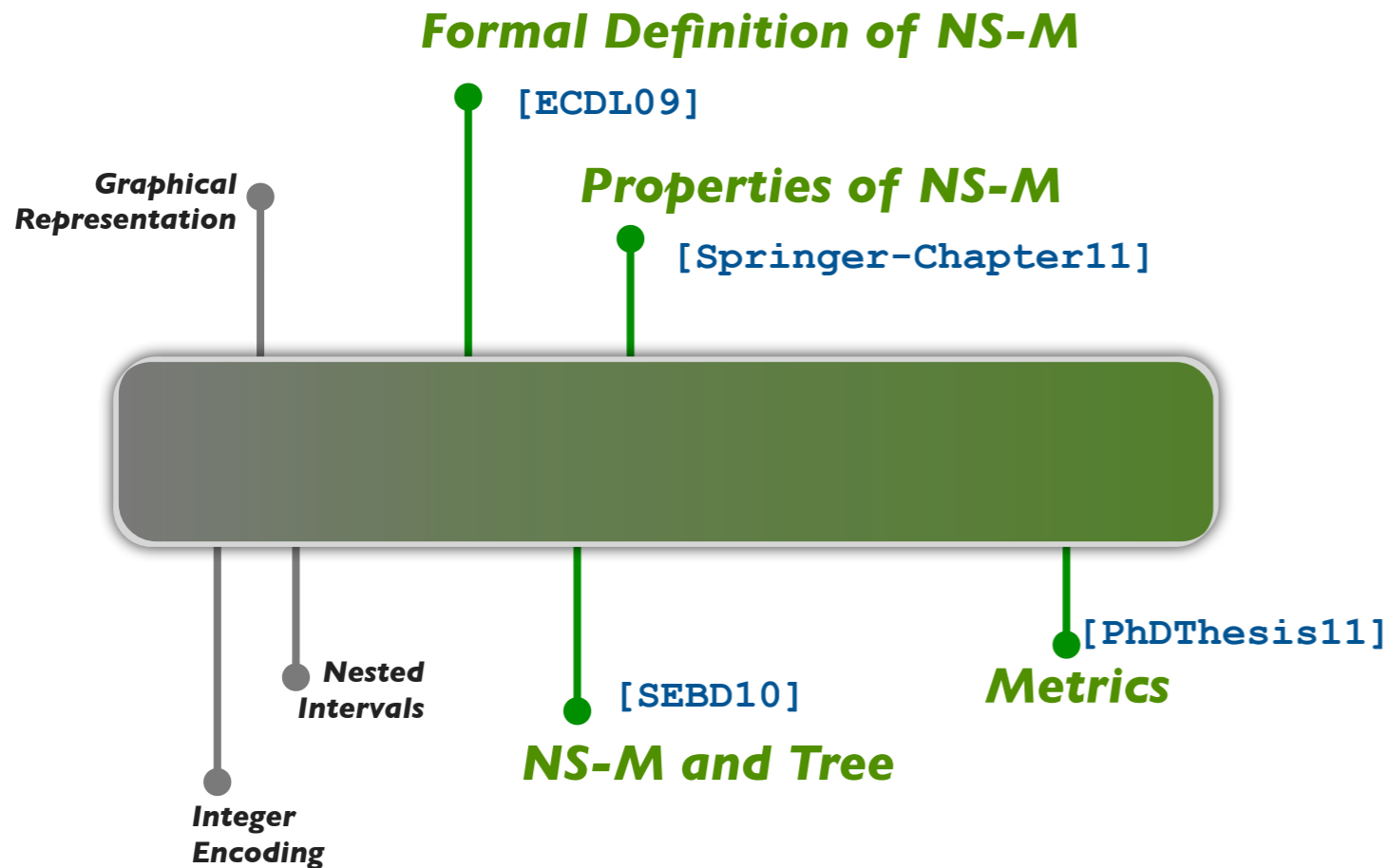
**Mapping Between the
Models**

Relationships with the Tree

Metrics

State of the art

Innovative



The NESTOR Model

The Nested Set Model
NS-M

The Inverse Nested Set Model
INS-M

Properties of the Set Data Models

Mapping Between the Models

Relationships with the Tree

Metrics

[ECDL09] N. Ferro and G. Silvello: "The NESTOR Framework: How to Handle Hierarchical Data Structures", Proc. of ECDL 2009 in LNCS 5741 series, pp. 215 - 226, Springer-Verlag, Germany, 2009.

[Springer-Chapter11] Agosti, M., Ferro, N. and Silvello, G. "How to Handle Hierarchically Structured Resources Addressing Interoperability Issues in Digital Libraries". In Learning Structure and Schemas from Documents, Studies in Computational Intelligence. M. Biba and Khafa, F. eds., Springer-Verlag, Germany 2011. In print.

[SEBD10] Agosti, A., Ferro, N., and Silvello, G. "The NESTOR Framework: Manage, Access and Exchange Hierarchical Data Structures". In Proceedings of SEBD 2010, pages 242-253. Italy, 2010.

[PhDThesis11] Silvello, G. "A Set-Based Approach to Deal with Hierarchical Structures", PhD Thesis, University of Padua, 2011.



The NESTOR Model

State of the art

Innovative

The NESTOR Model

The Nested Set Model
NS-M

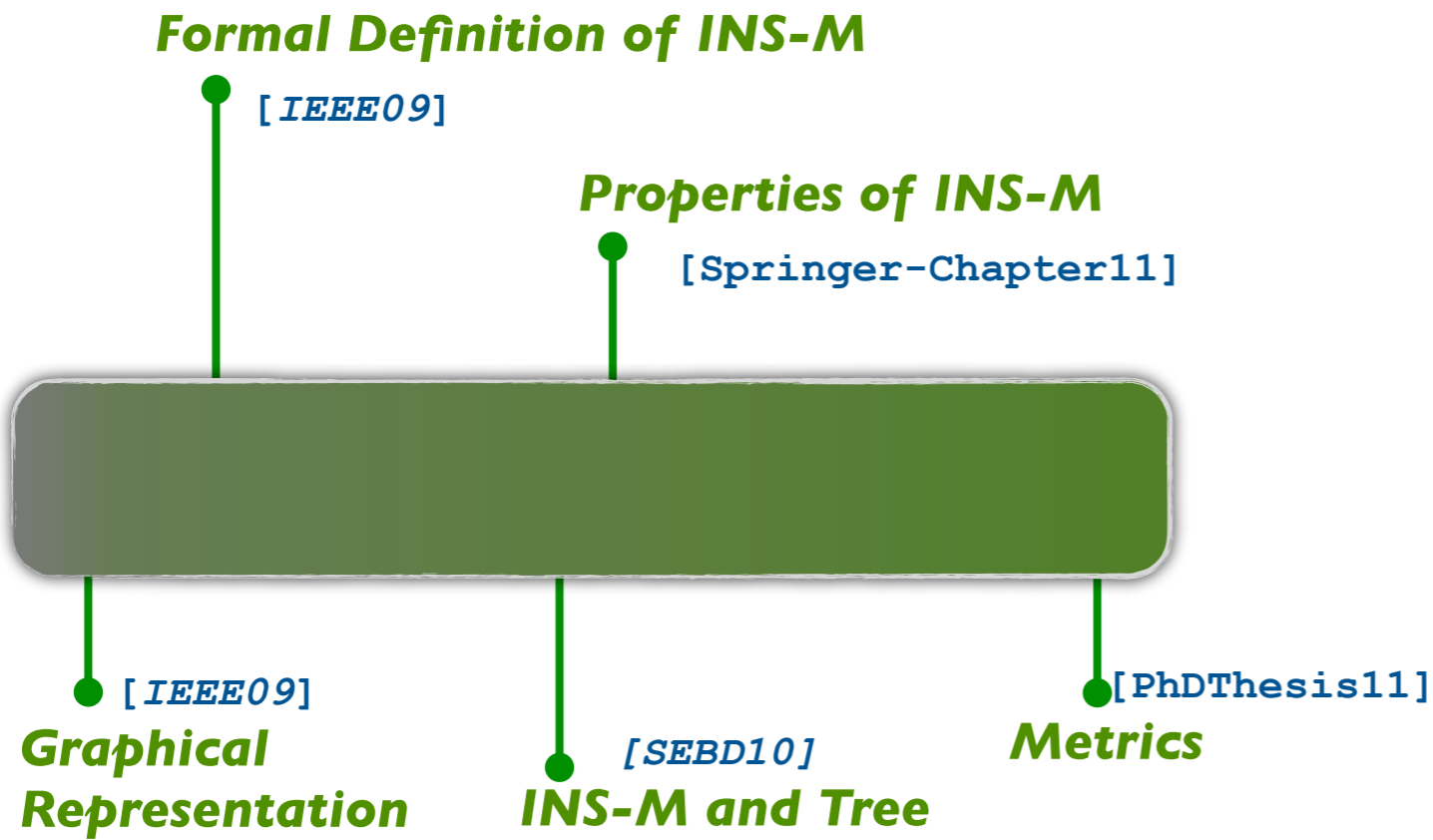
The Inverse Nested Set Model
INS-M

Properties of the Set Data Models

Mapping Between the Models

Relationships with the Tree

Metrics

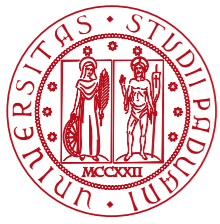


[IEEE09] Agosti, M., Ferro, N., and Silvello, G. Access and Exchange of Hierarchically Structured Resources on the Web with the NESTOR Framework. Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference, pages 659–662, 2009.

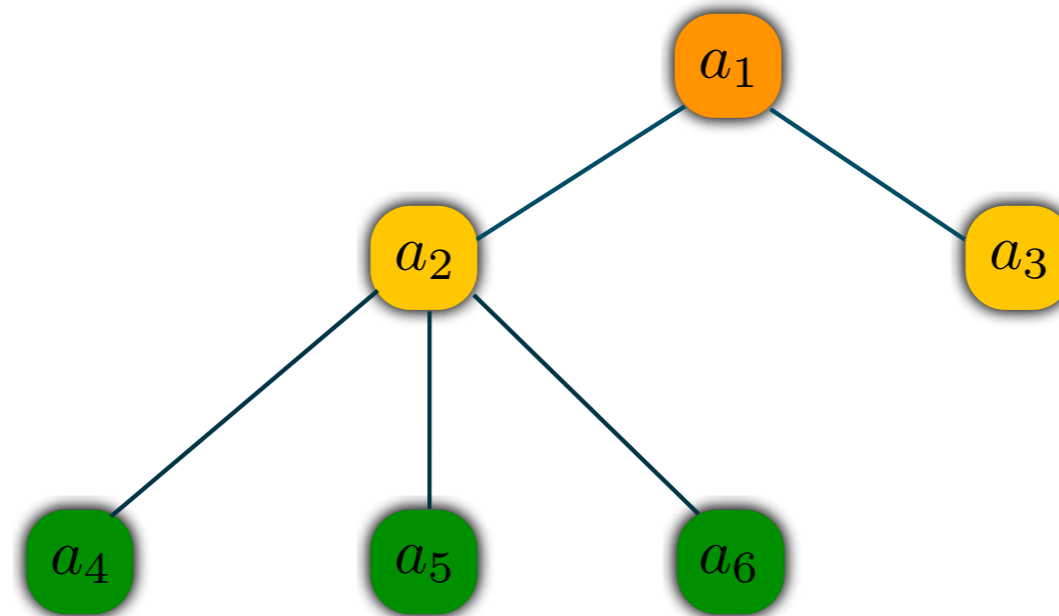
[Springer-Chapter11] Agosti, M., Ferro, N. and Silvello, G. "How to Handle Hierarchically Structured Resources Addressing Interoperability Issues in Digital Libraries". In Learning Structure and Schemas from Documents, Studies in Computational Intelligence. M. Biba and Khafa, F. eds., Springer-Verlag, Germany 2011. In print.

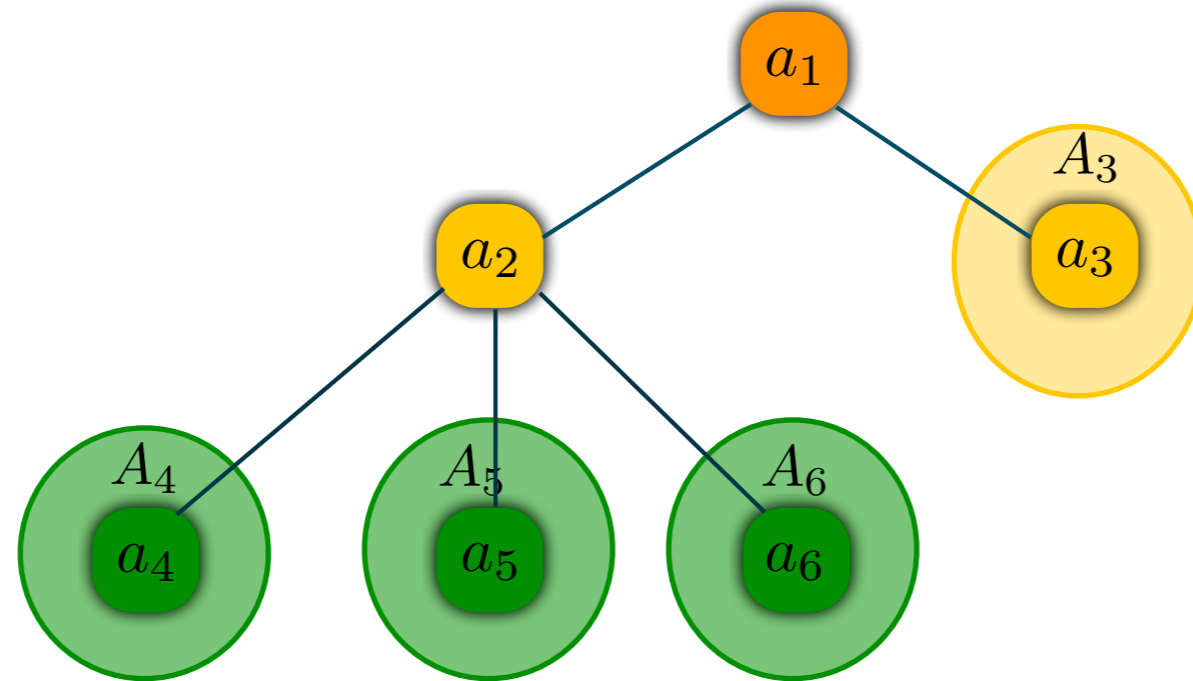
[SEBD10] Agosti, A., Ferro, N., and Silvello, G. The NESTOR Framework: Manage, Access and Exchange Hierarchical Data Structures. In Proceedings of SEBD 2010, pages 242–253. Italy, 2010.

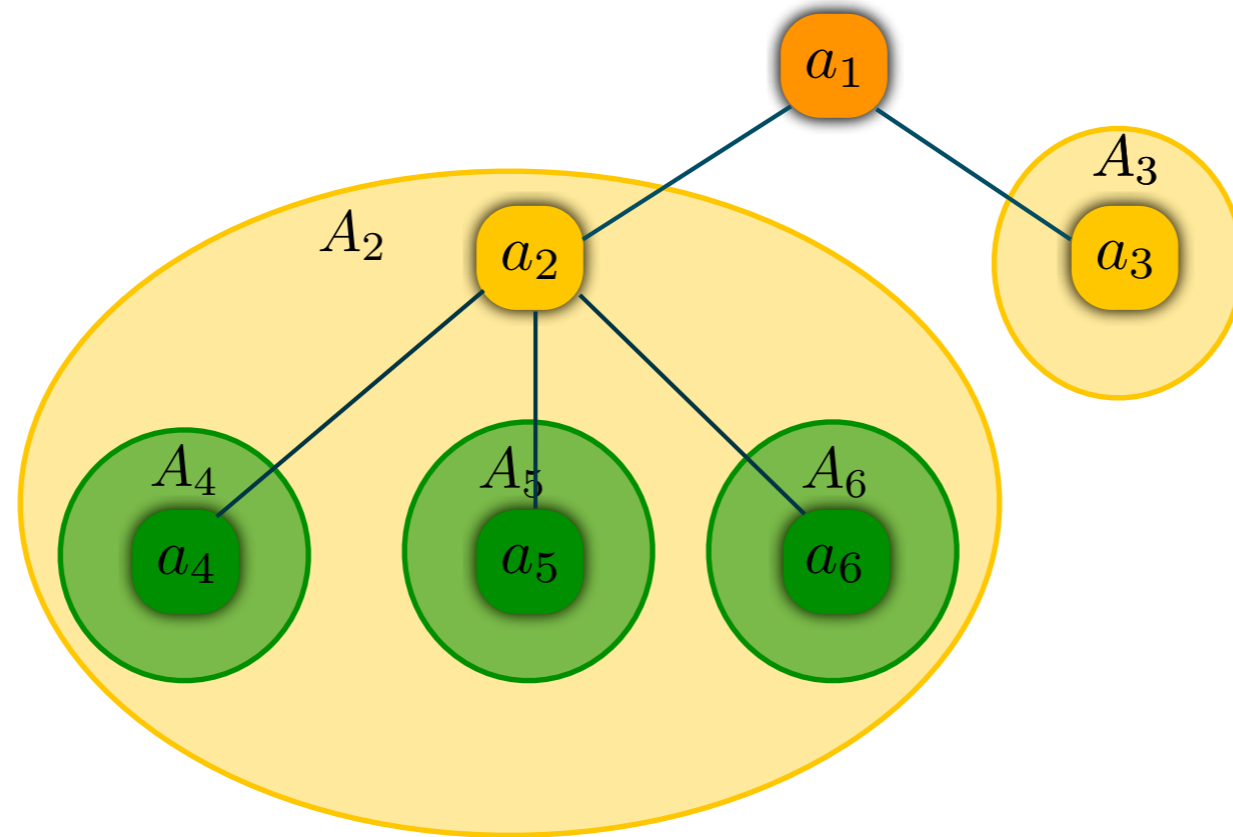
[PhDThesis11] Silvello, G. "A Set-Based Approach to Deal with Hierarchical Structures", PhD Thesis, University of Padua, 2011.

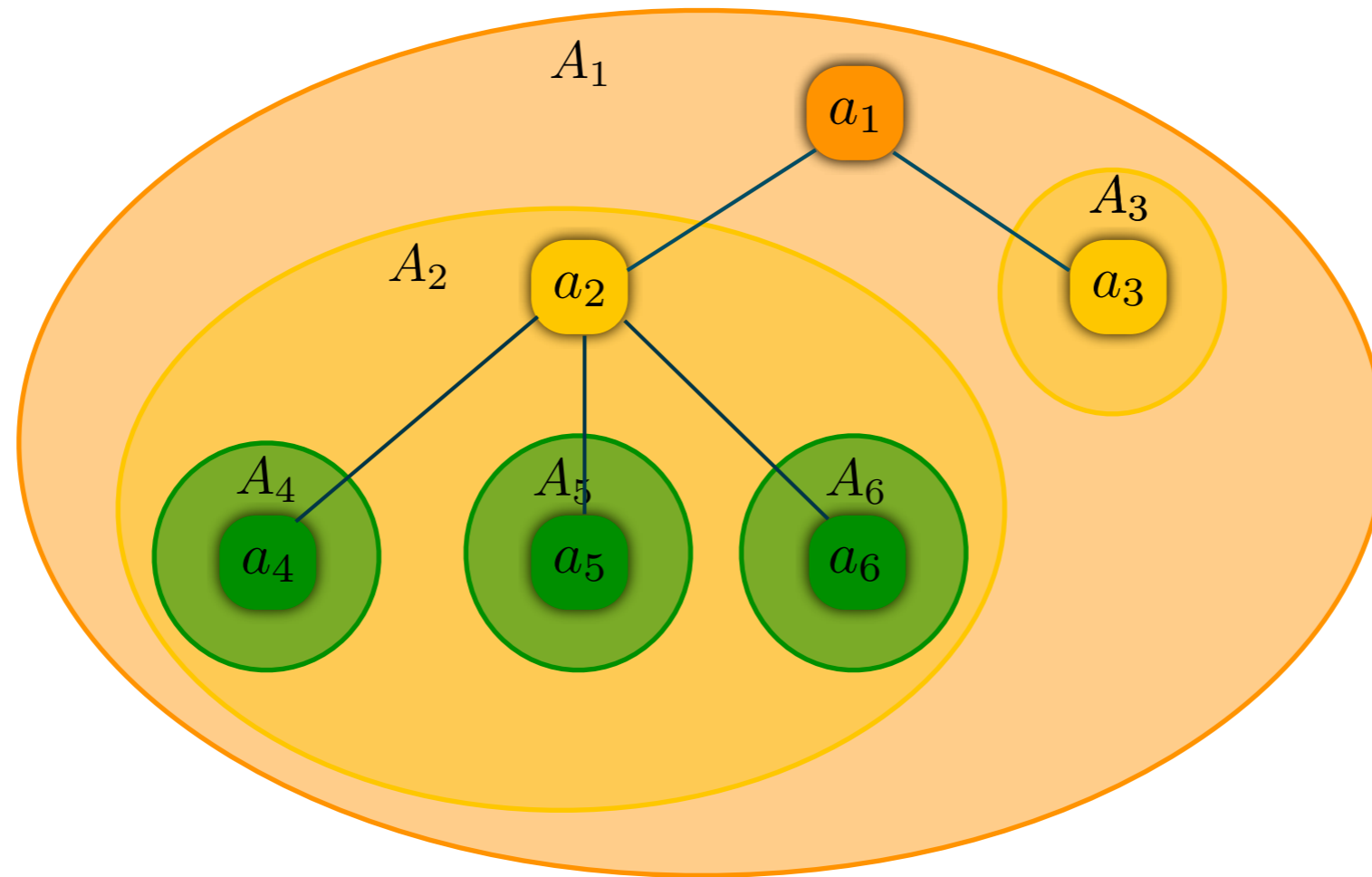


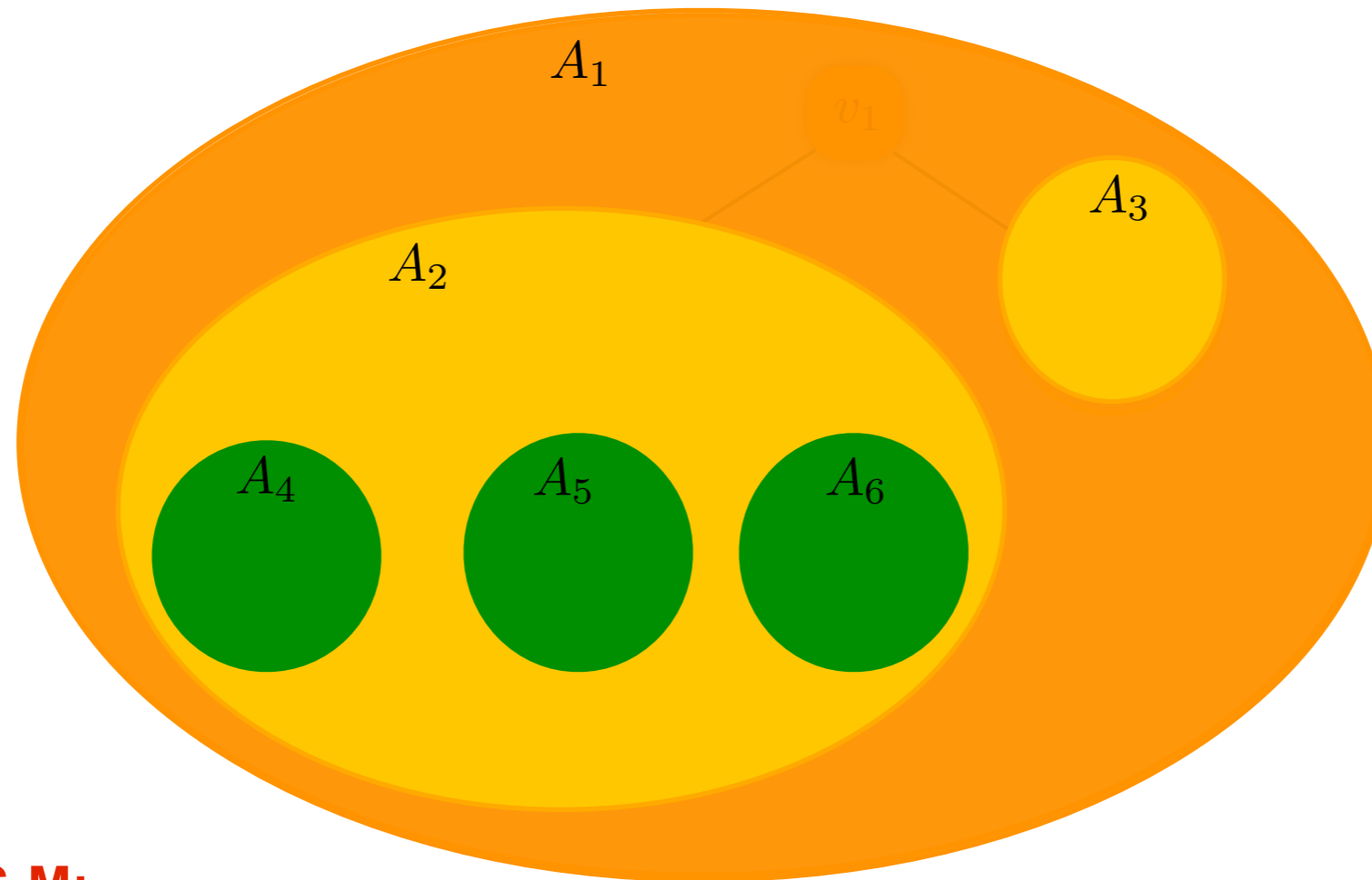
The Nested Set Model











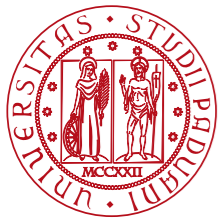
Definition of NS-M:

Let A be a set and let $\{A_i\}_{i \in I}$ be a family. Then $\{A_i\}_{i \in I}$ is a **Nested Set Family** if:

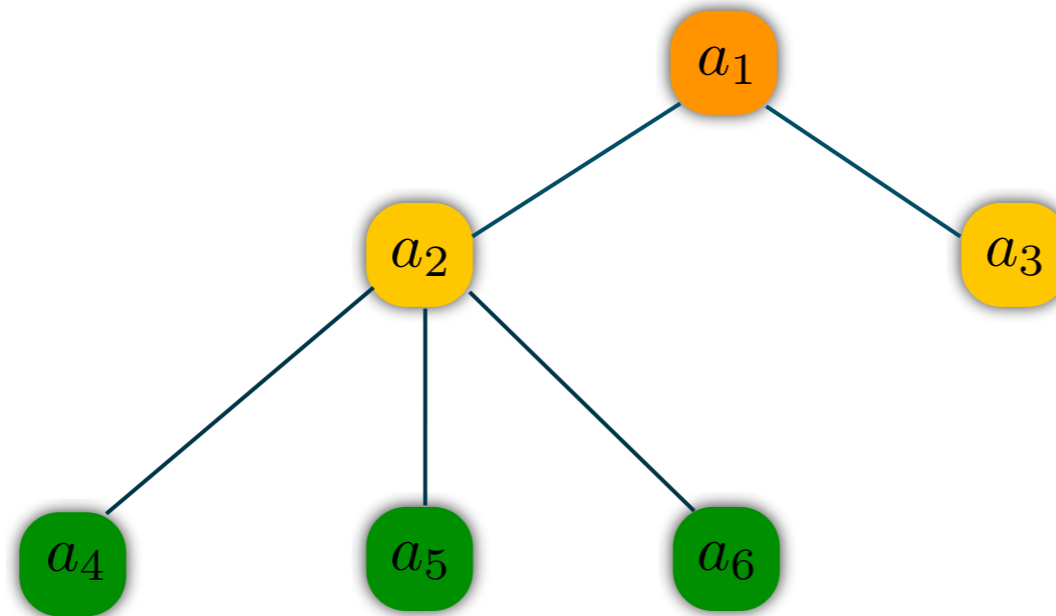
$$A \in \{A_i\}_{i \in I}$$

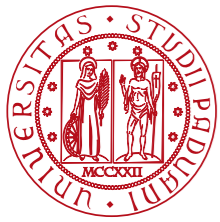
$$\emptyset \notin \{A_i\}_{i \in I}$$

$$\forall A_h, A_k \in \{A_i\}_{i \in I}, h \neq k \mid A_h \cap A_k \neq \emptyset \Rightarrow A_h \subset A_k \vee A_k \subset A_h$$

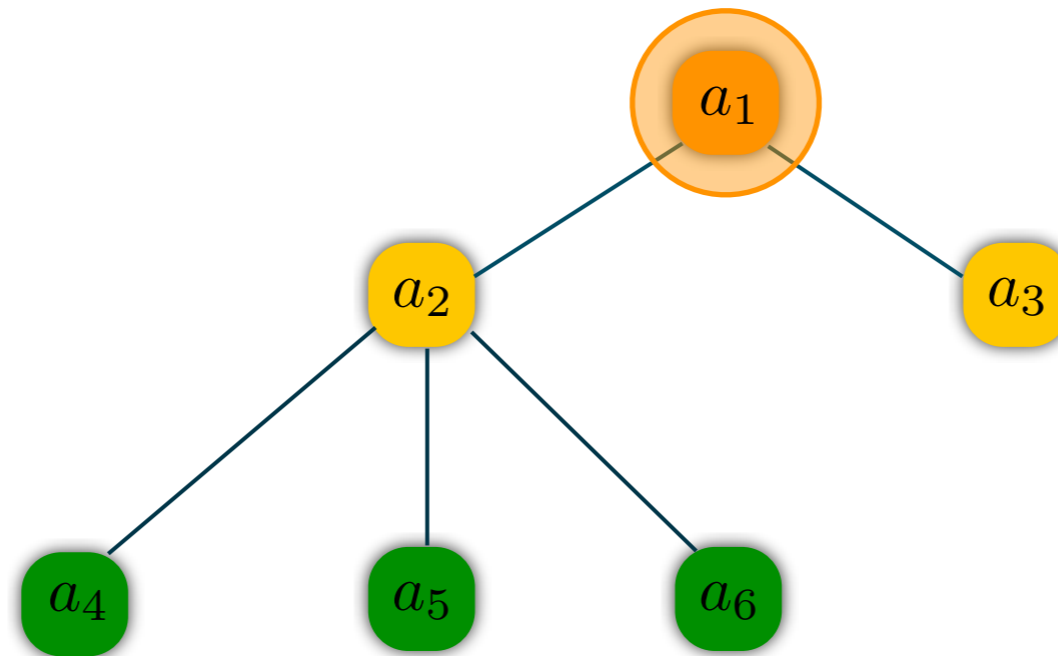


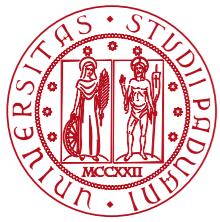
The Inverse Nested Set Model



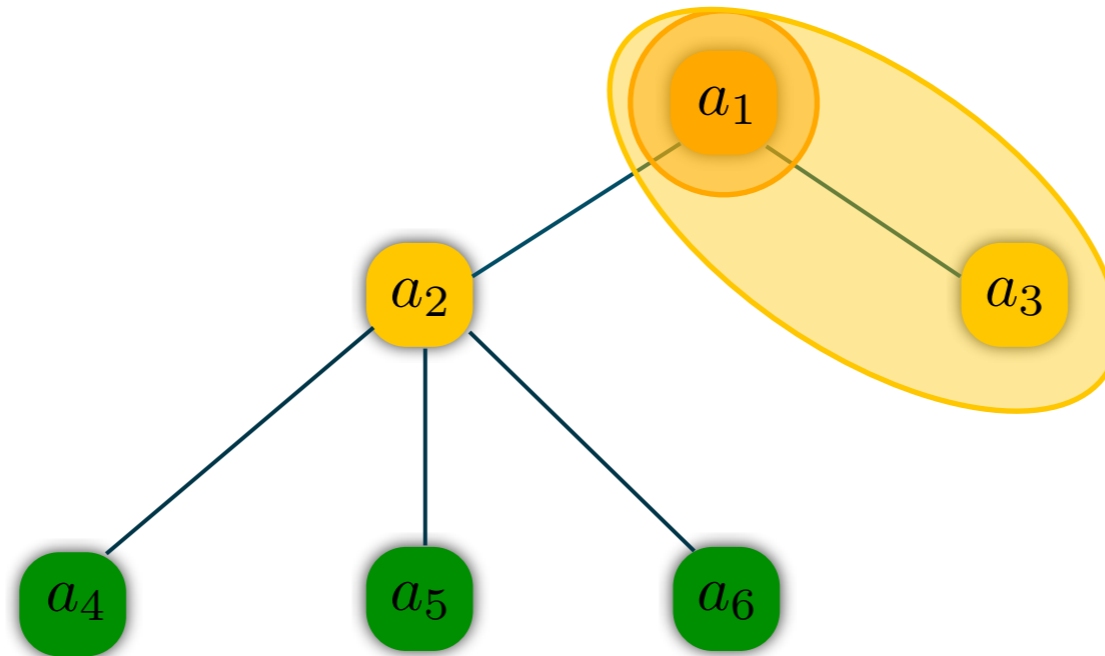


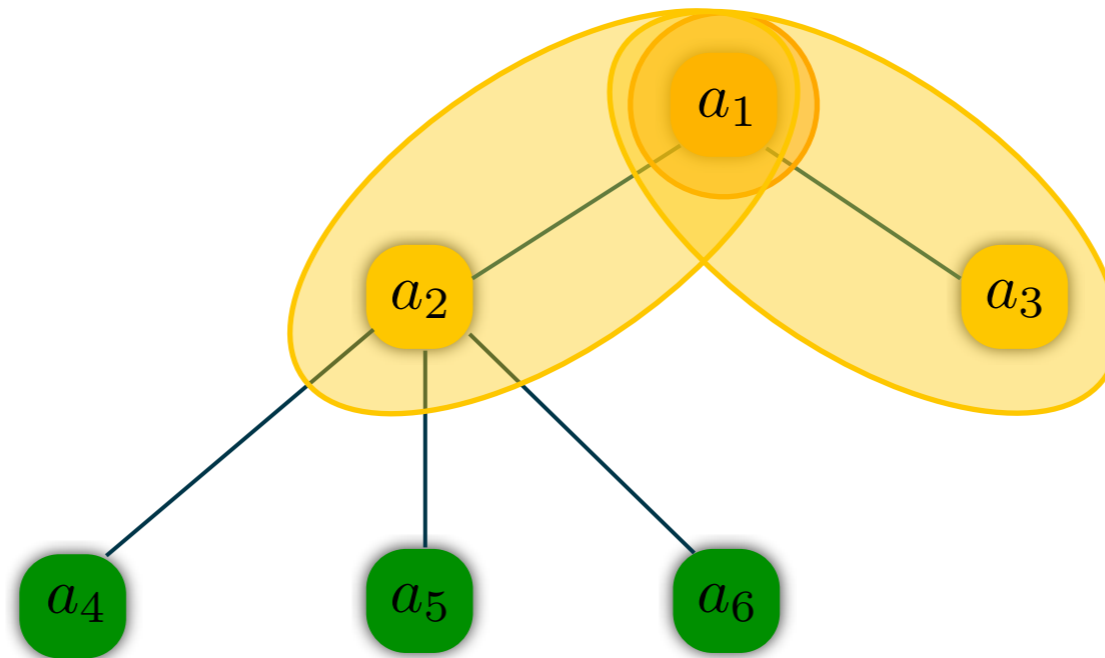
The Inverse Nested Set Model

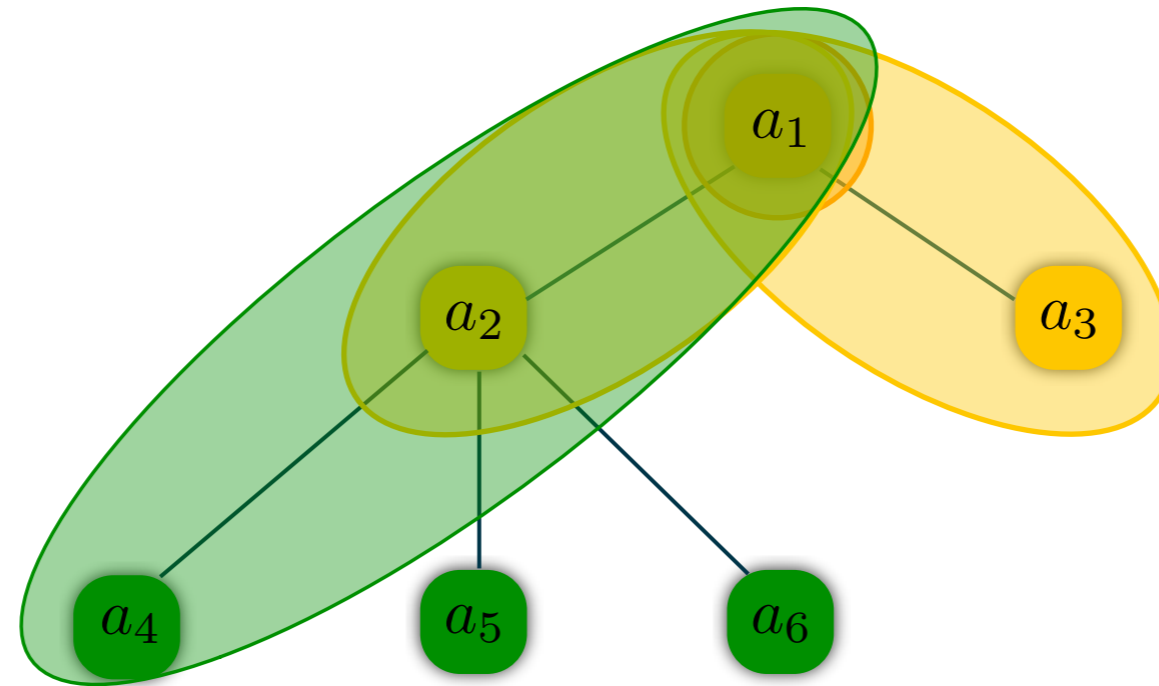


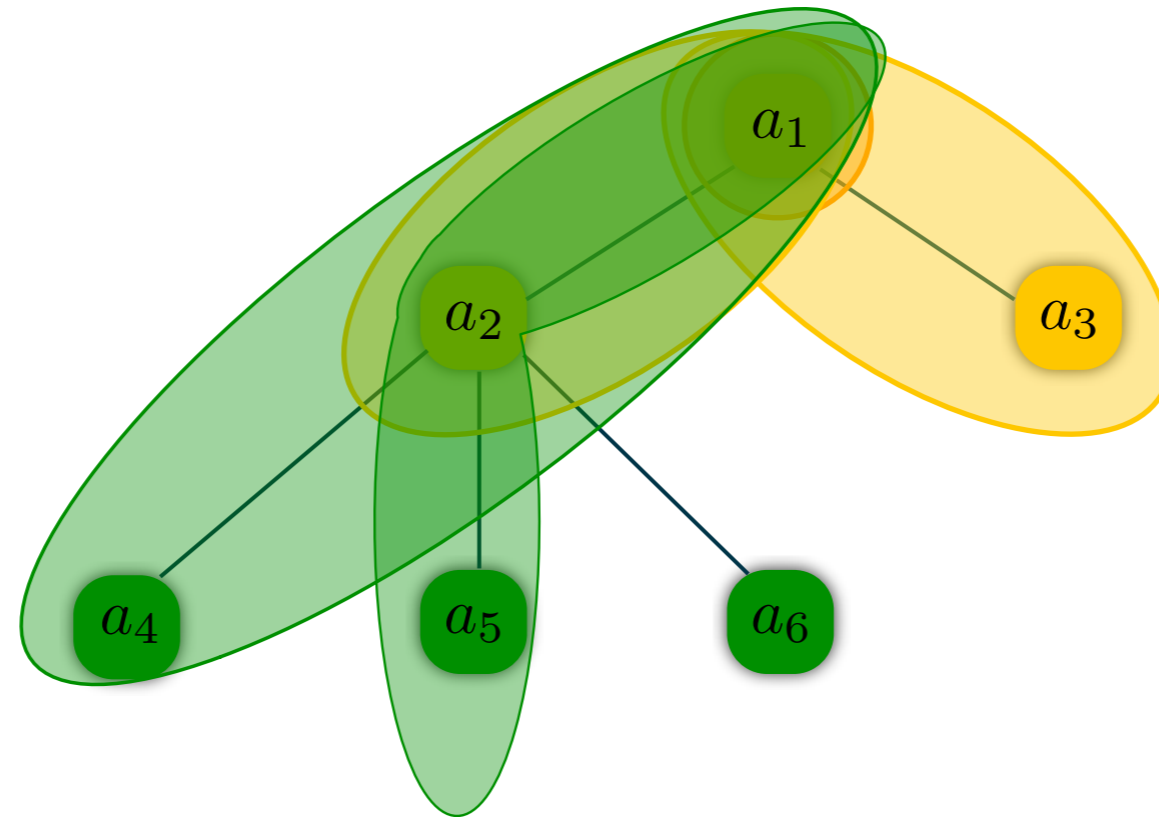


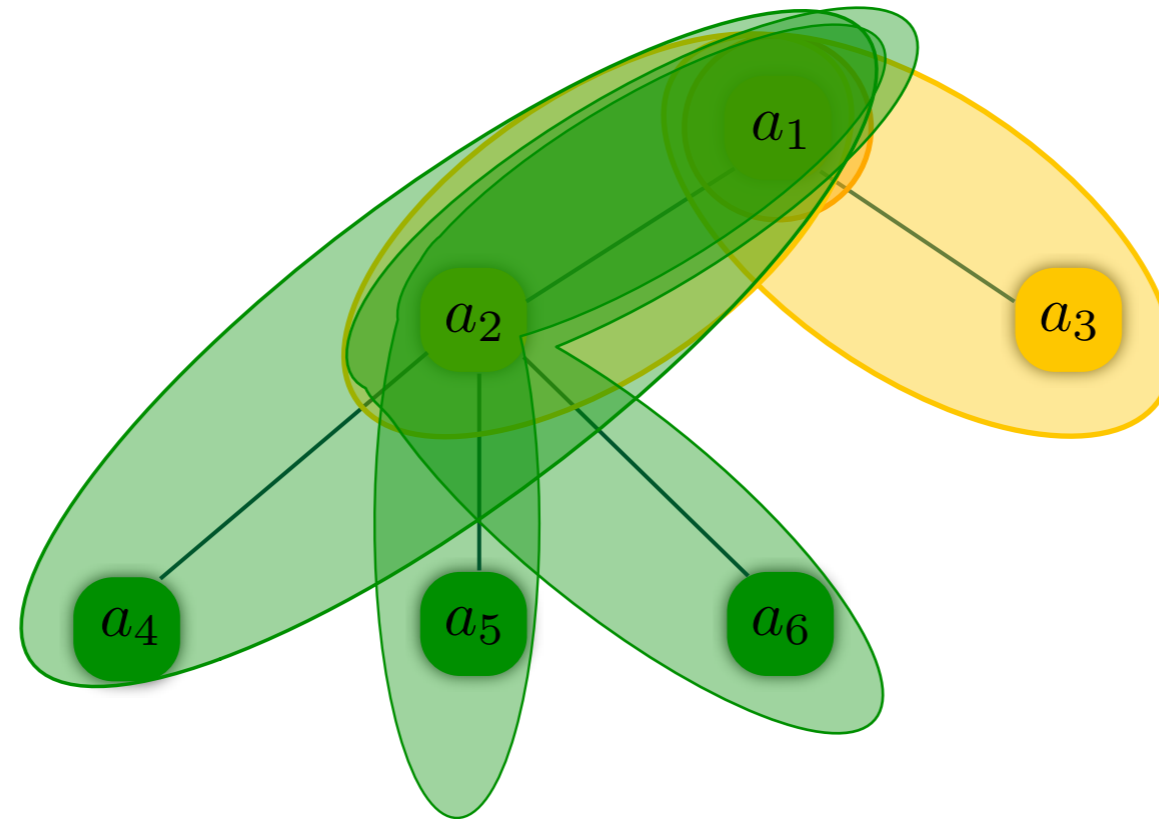
The Inverse Nested Set Model





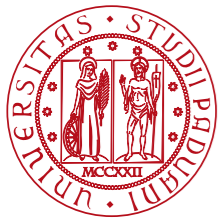




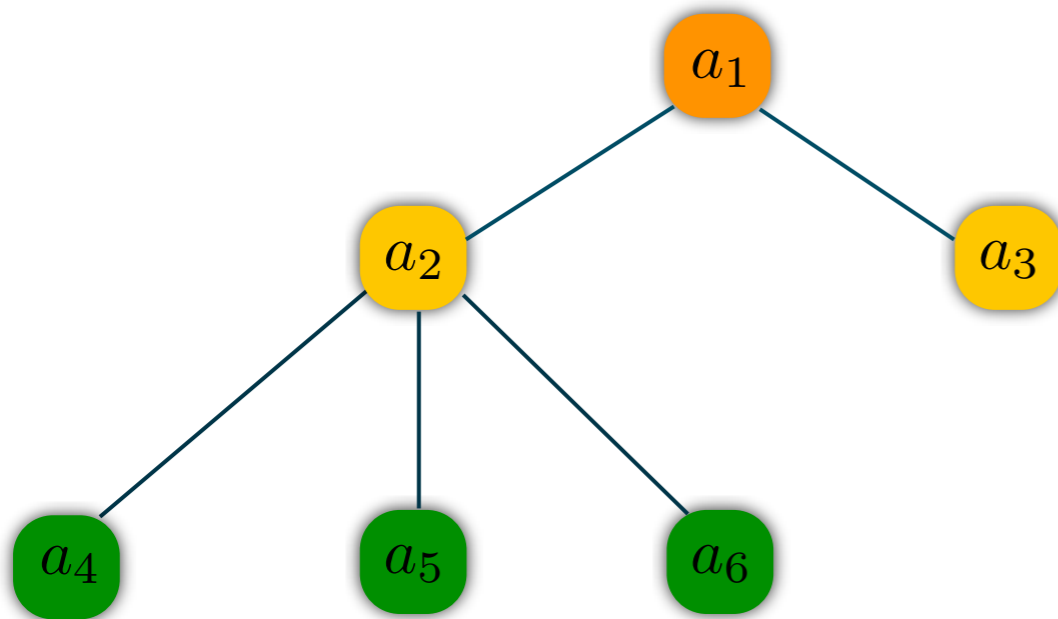


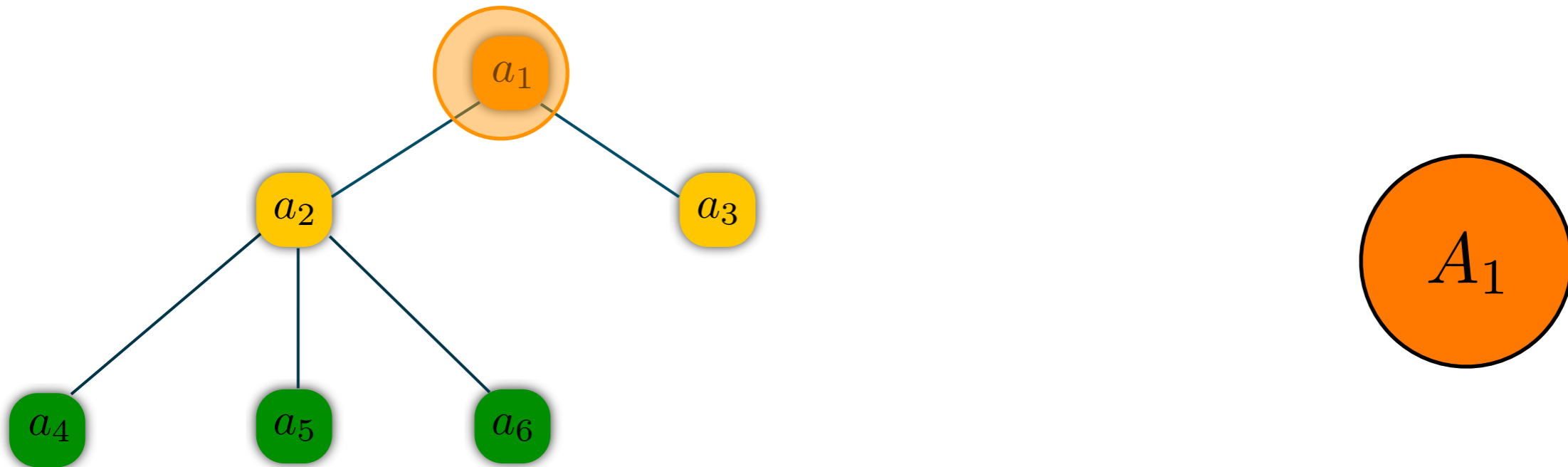


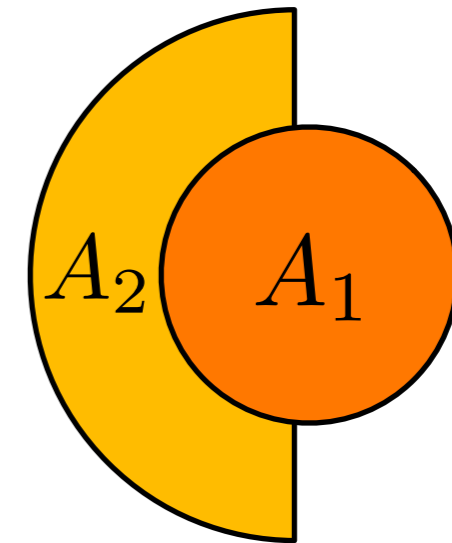
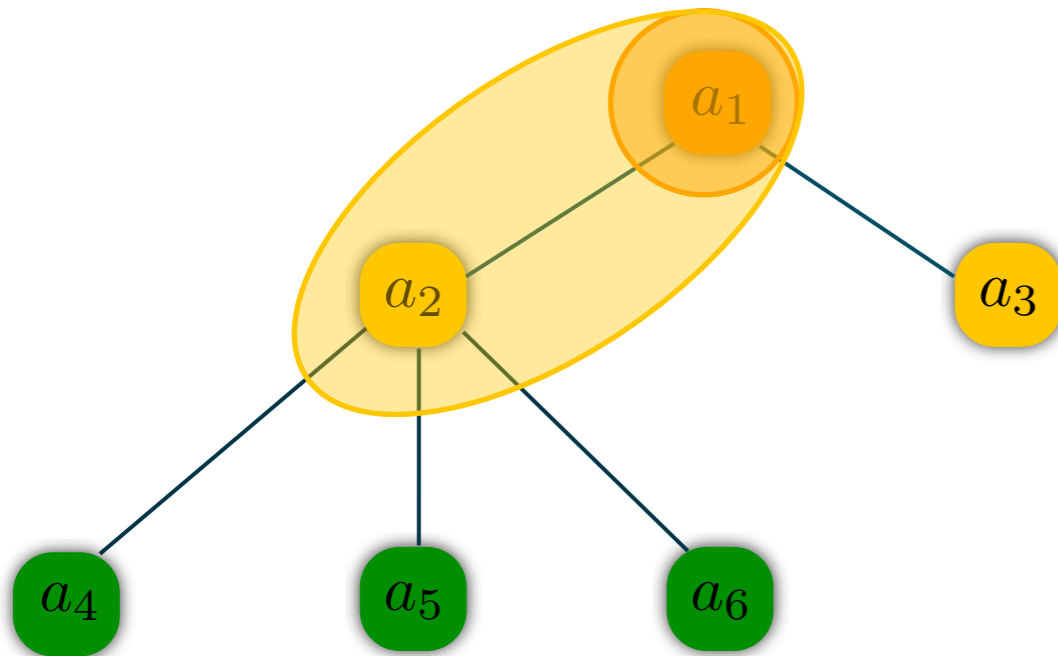
The Inverse Nested Set Model

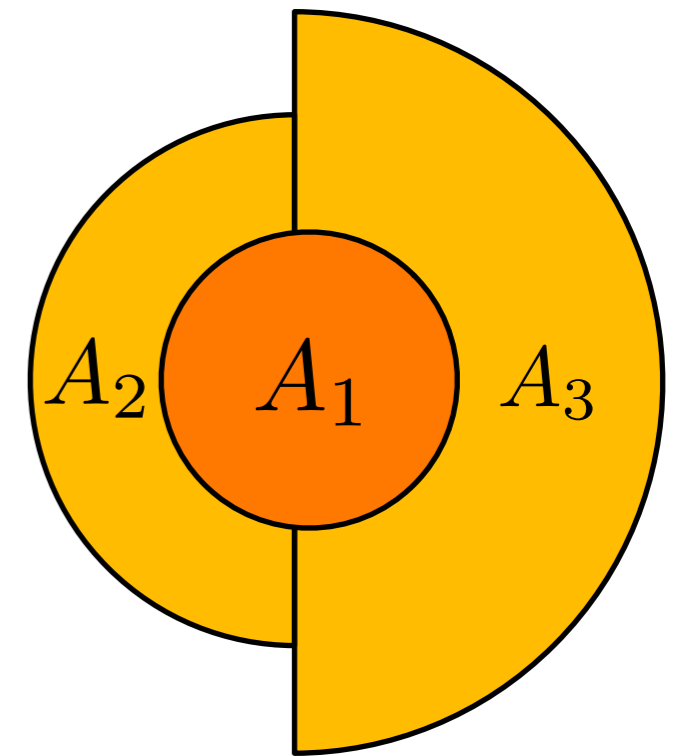
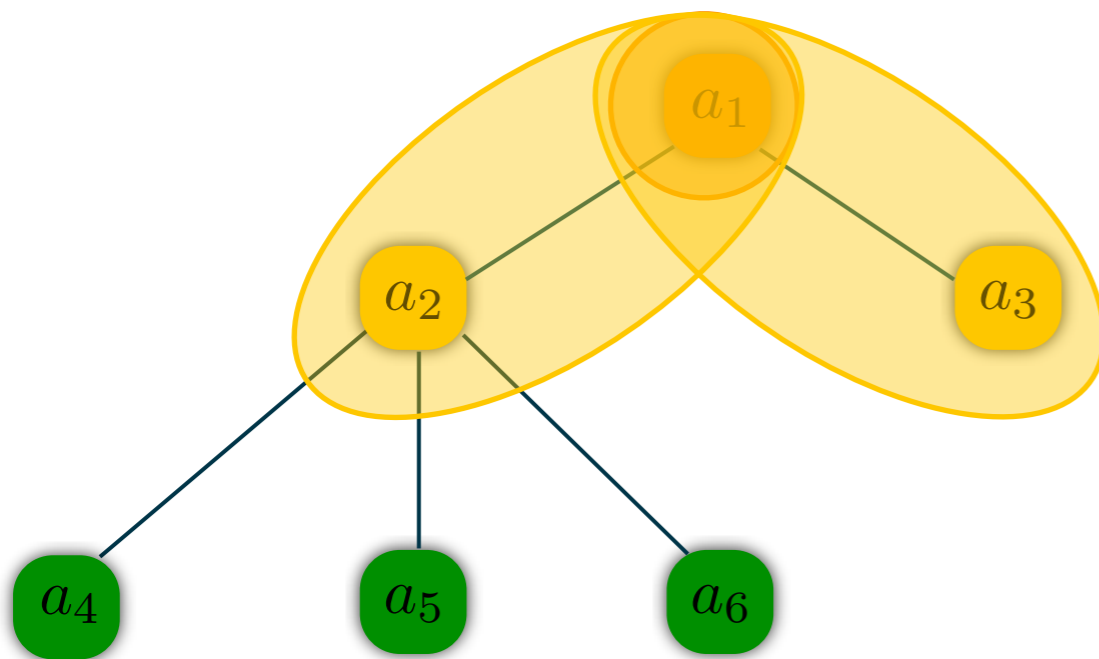


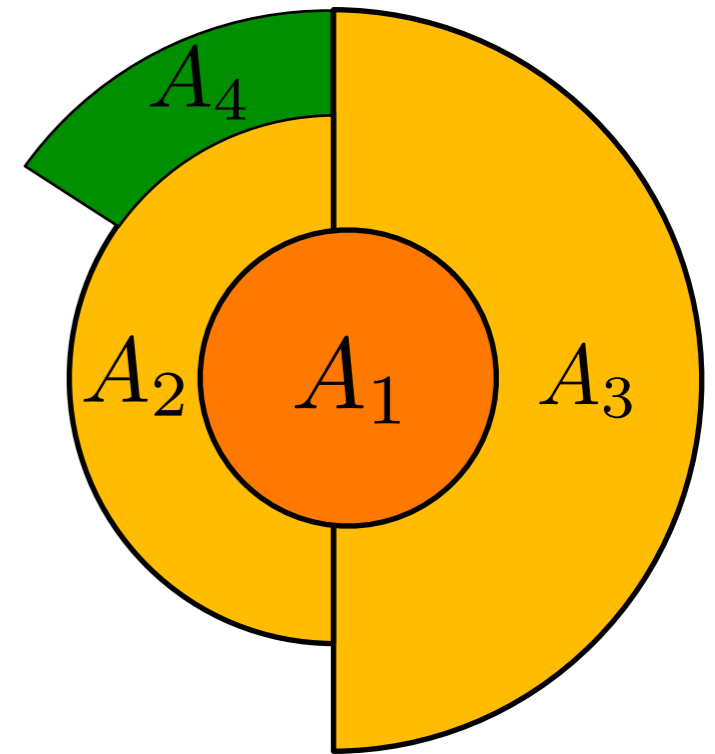
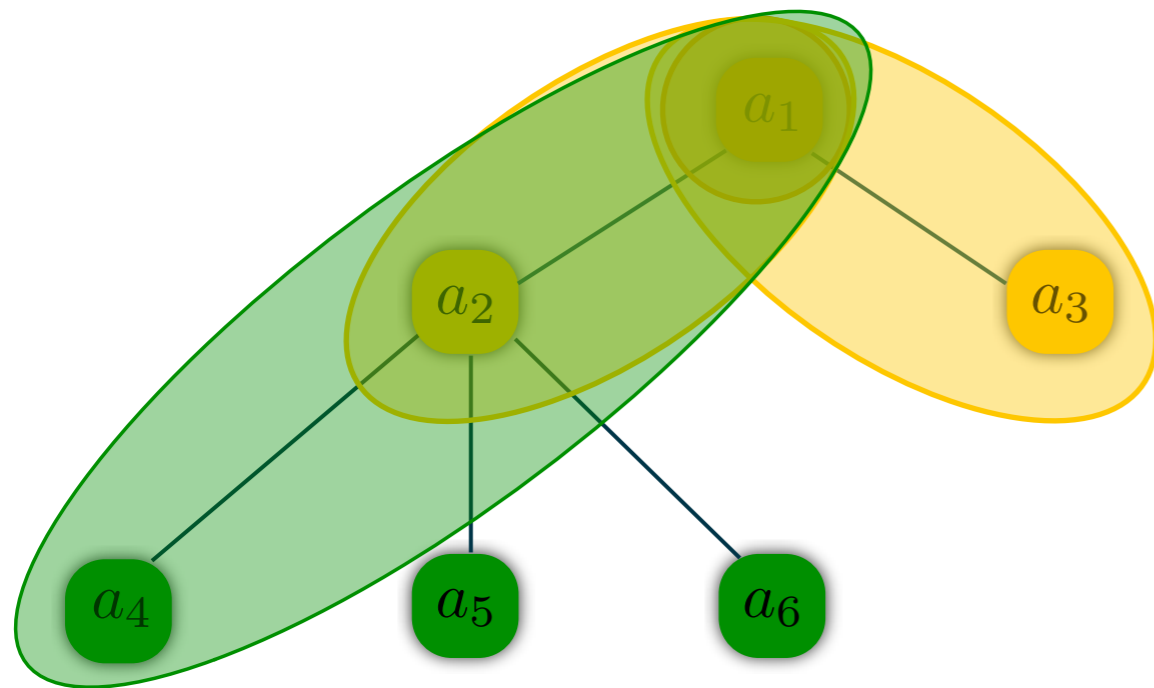
The Inverse Nested Set Model

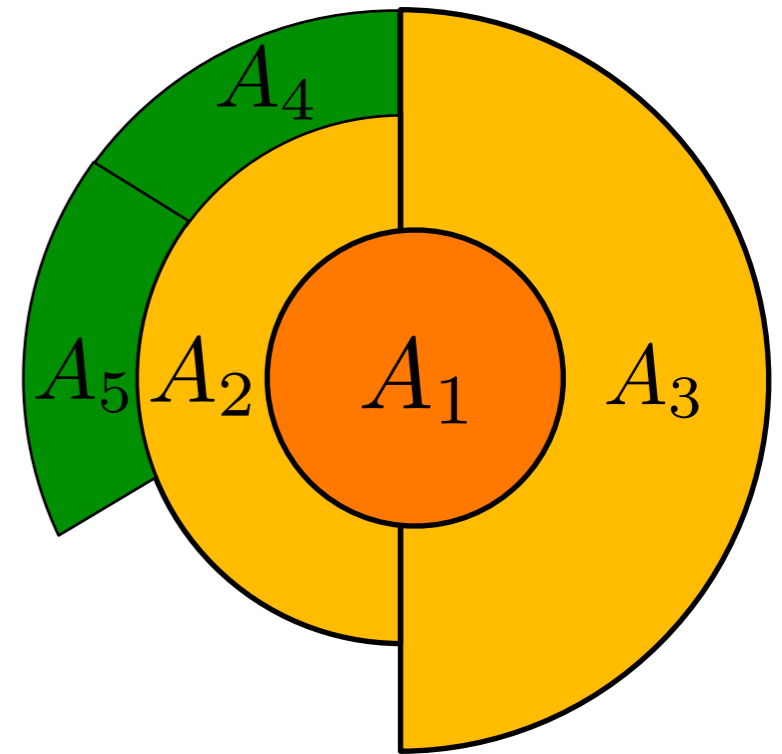
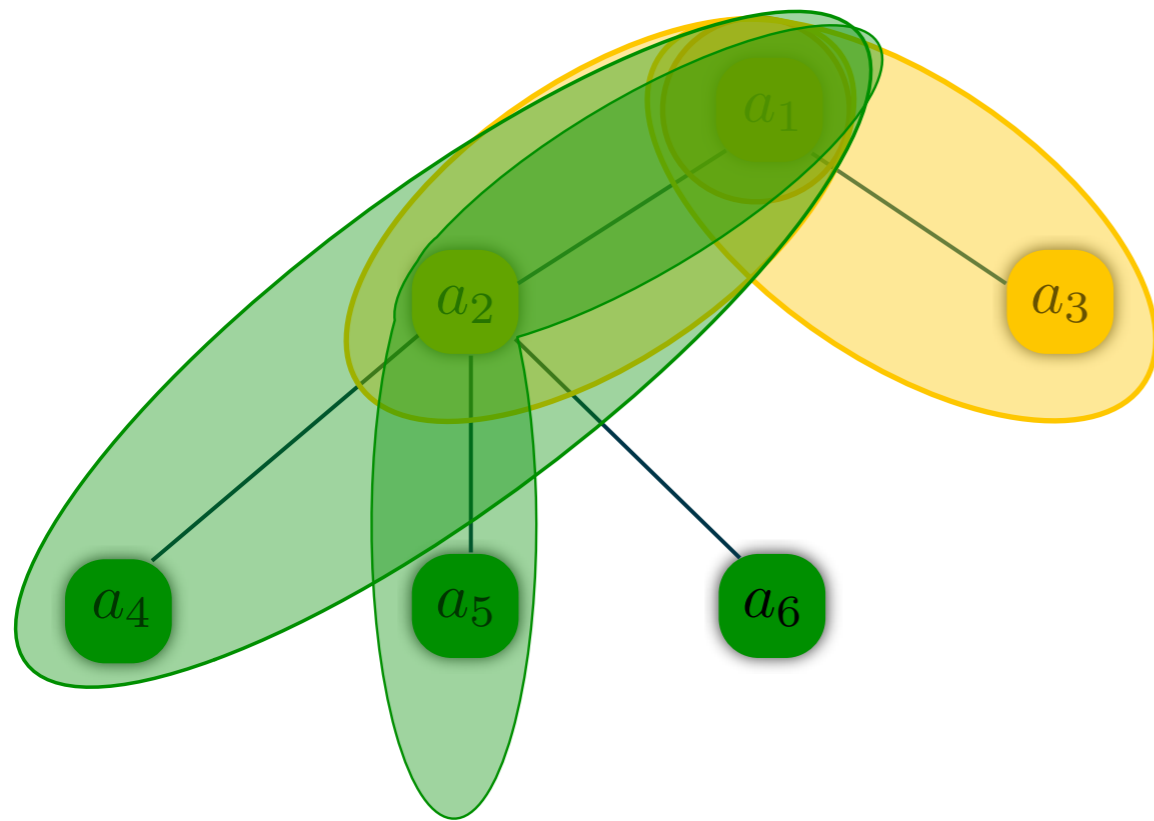


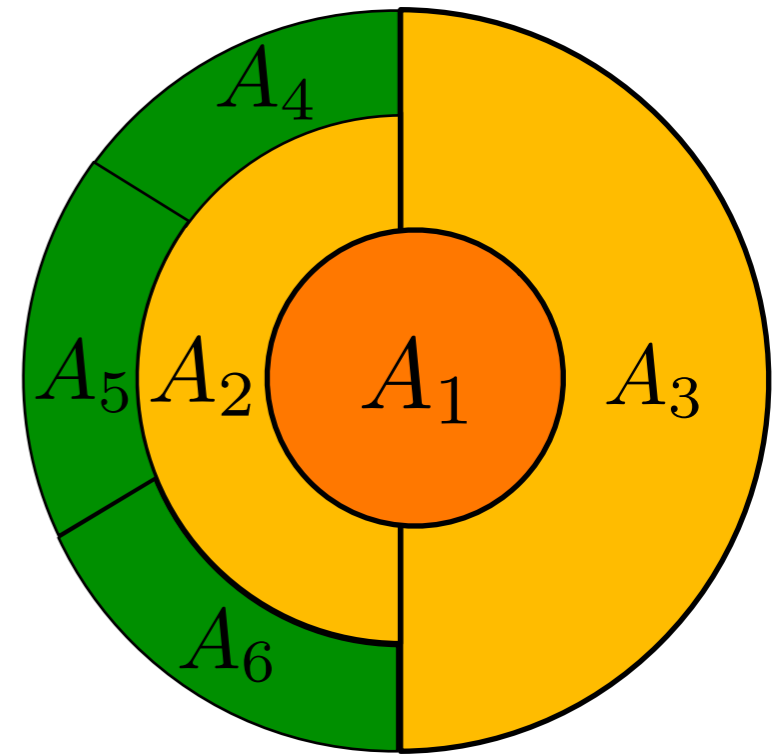
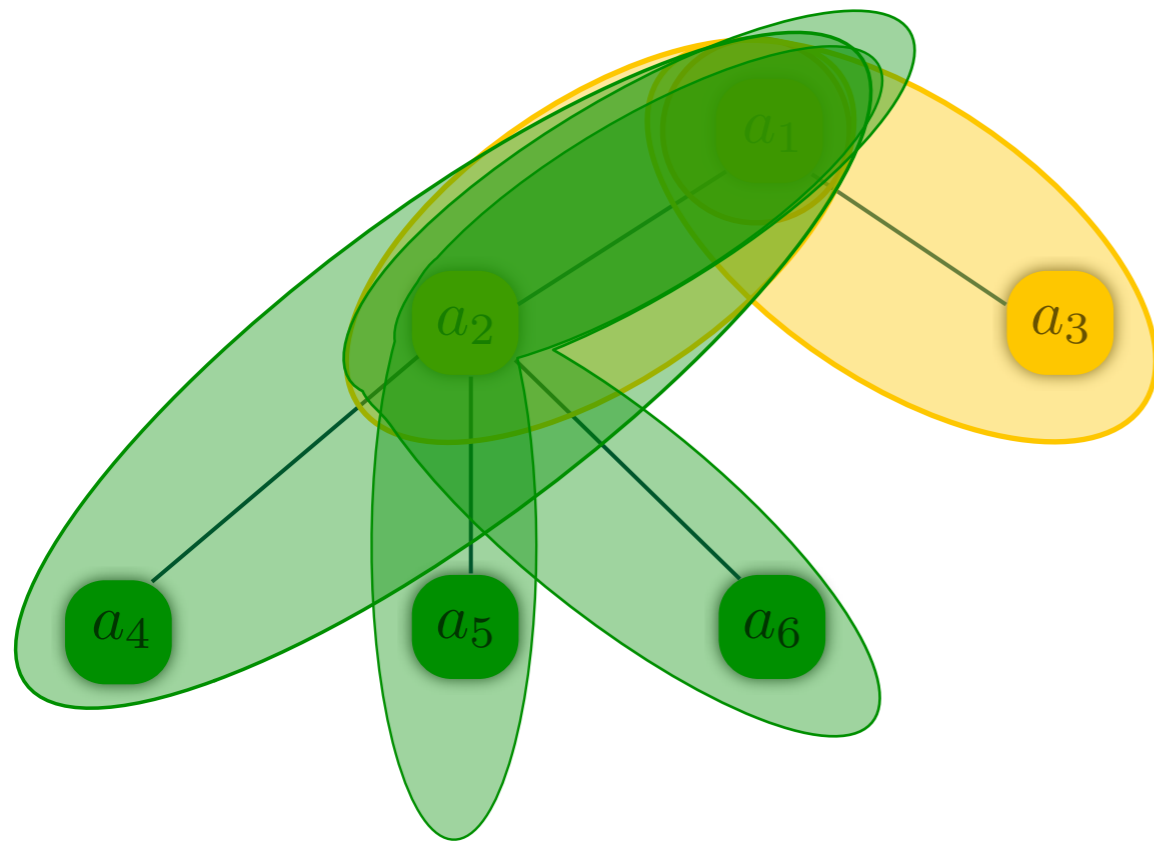


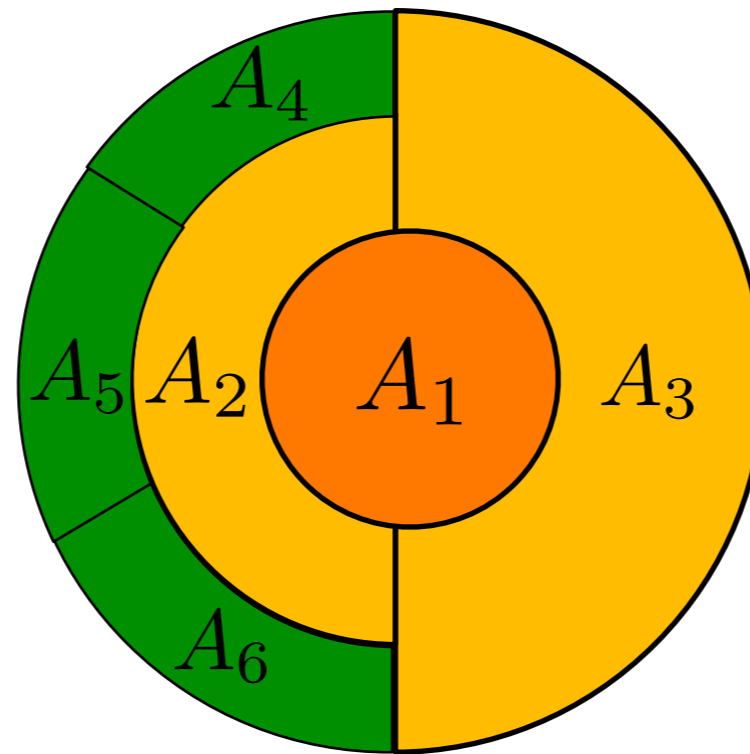












Definition of INS-M:

Let A be a set and let $\{A_i\}_{i \in I}$ be a family. Then $\{A_i\}_{i \in I}$ is an **Inverse Nested Set Family** if:

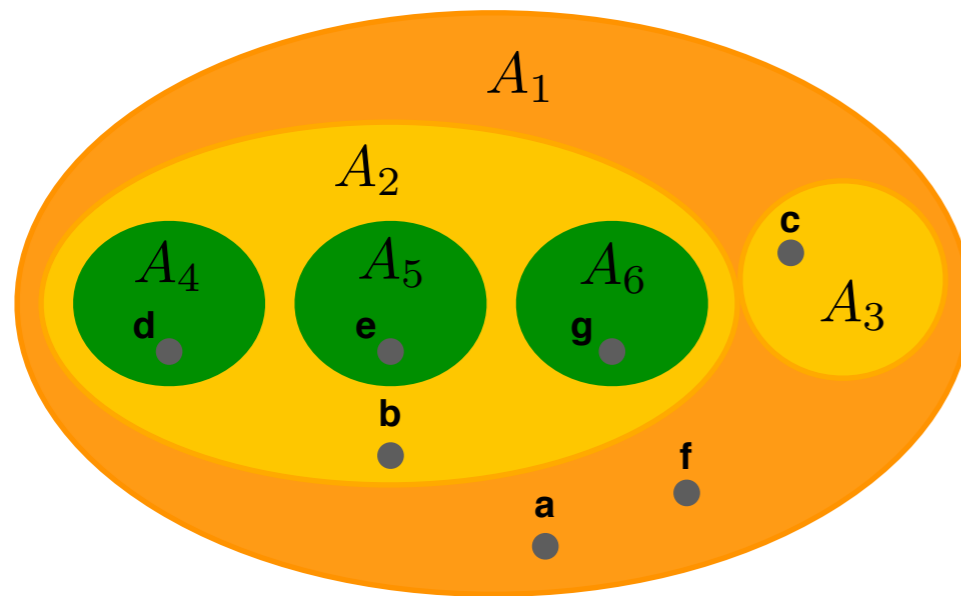
$$\emptyset \notin \{A_i\}_{i \in I}$$

$$\forall \{B_j\}_{j \in J} \subseteq \{A_i\}_{i \in I} \Rightarrow \bigcap_{j \in J} B_j \in \{A_i\}_{i \in I}.$$

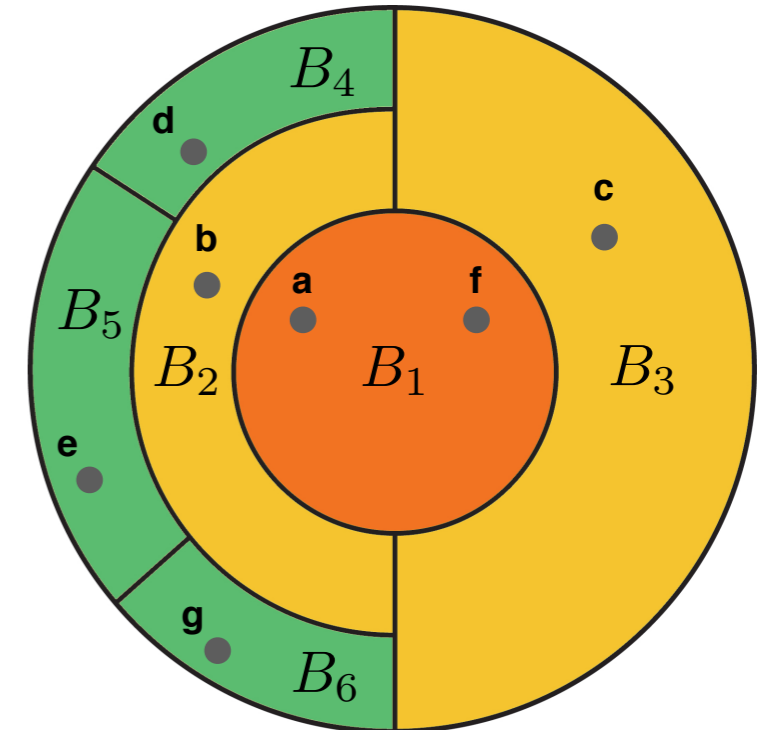
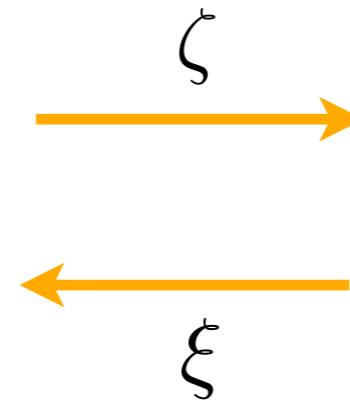
$$\forall \{B_j\}_{j \in J} \subseteq \{A_i\}_{i \in I}$$

$$\Rightarrow \exists B_k \in \{B_j\}_{j \in J} \mid \forall B_h \in \{B_j\}_{j \in J}, B_h \subseteq B_k$$

$$\Rightarrow \forall B_h, B_g \in \{B_j\}_{j \in J}, B_h \subseteq B_g \vee B_g \subseteq B_h.$$



$\{A_i\}_{i \in I}$



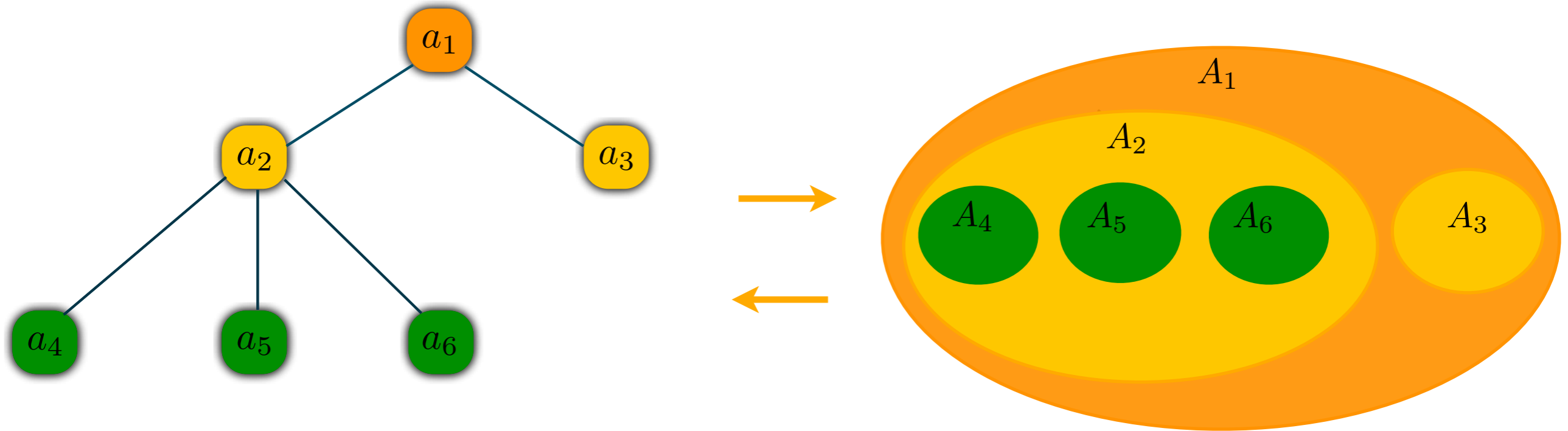
$\{B_j\}_{j \in J}$

Function ζ : From NS-M to INS-M

Let $\{A_i\}_{i \in I}$ be a family of sets. We define $\zeta : \{A_i\}_{i \in I} \rightarrow \{B_j\}_{j \in J}$ to be a function such that $\forall A_k \in \{A_i\}_{i \in I}, \exists B_k \in \{B_j\}_{j \in J} \mid B_k = \bigcup_{A_t \in \{A_i\}_{i \in I}} (A_t \setminus \bigcup_{A_s \in \mathcal{A}} S_{\mathcal{A}}^+(A_s))$.

Function ξ : From INS-M to NS-M

Let $\{A_i\}_{i \in I}$ be a family of sets. We define $\xi : \{A_i\}_{i \in I} \rightarrow \{B_j\}_{j \in J}$ to be a function such that $\forall A_k \in \{A_i\}_{i \in I}, \exists B_k \in \{B_j\}_{j \in J} \mid B_k = \bigcup_{A_t \in \{A_i\}_{i \in I}} (A_t \cup S_{\mathcal{A}}^-(A_t)) \setminus \bigcup_{A_s \in \mathcal{A}} S_{\mathcal{A}}^+(A_s)$.

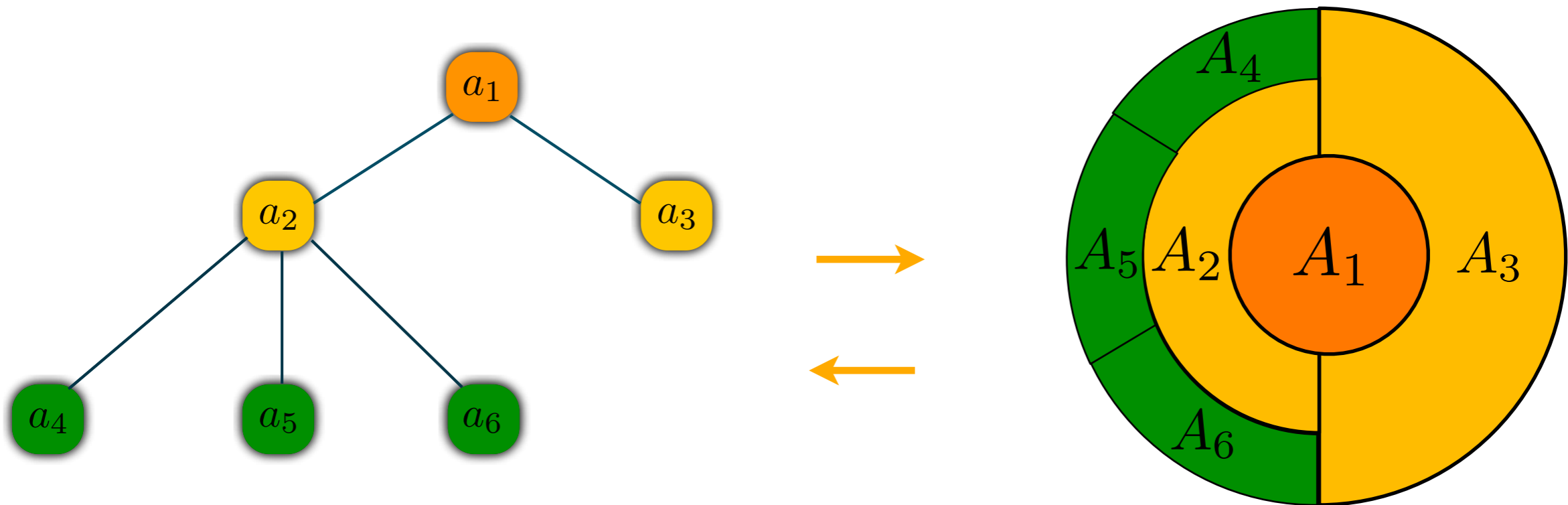


Theorem: From Tree to NS-M

Let \mathcal{V}_V be a NS-F, V be a set of nodes and E be a set of edges where $\forall v_j \in V, \exists! V_{v_j} \in \mathcal{V}_V \wedge \forall e_{j,k} \in E, \exists! V_{v_j}, A_k \in \mathcal{V}_V \mid A_k \subset V_{v_j}$. Then $T = (V, E)$ is a tree.

Theorem: From NS-M to Tree

Let $T = (V, E)$ be a tree and let \mathcal{V}_V be a family where the set of nodes V is its index set of the family and $\forall v_i \in V, V_{v_i} = \Gamma^+(v_i)$. Then \mathcal{V}_V is a Nested Set Family.

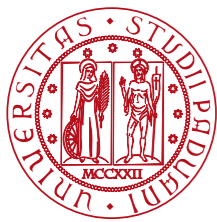


Theorem: From Tree to INS-M

Let $T = (V, E)$ be a tree and let \mathcal{V}_V be a family where the set of nodes V is its index set of the family and $\forall v_i \in V, V_{v_i} = \Gamma^-(v_i)$. Then \mathcal{V}_V is an Inverse Nested Set family.

Theorem: From INS-M to Tree

Let \mathcal{V}_V be a INS-F, V be a set of nodes and E be a set of edges where $\forall v_j \in V, \exists! V_{v_j} \in \mathcal{V}_V \wedge \forall e_{j,k} \in E, \exists! V_{v_j}, V_{v_k} \in \mathcal{V}_V \mid V_{v_j} \subset V_{v_k}$. Then $T = (V, E)$ is a tree.



- **Distance Between Sets:**
 - **Graphical Distance** [Diestel106]: correspondence with graphical distance in trees.
- **Distance Between Families of Subsets:**
 - **Content-Based** [Jaccard1901].
 - **Structure-Based** [Galle10,ZezulaEtAl106].
 - **NESTOR Distance:** Weighted Linear Combination of Content- and Structure-Based Distances.

[Diestel106] Diestel, R. "Graph Theory". Springer, Berlin Heidelberg, Germany, 2006.

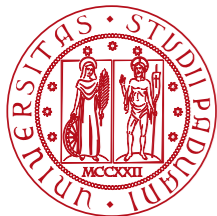
[Jaccard1901] Jaccard, P. "Etude comparative de la distribution florale dans une portion des alpes et des jura". Bulletin del la Societè Vaudoise des Sciences Naturelles, 37:547-579, 1901.

[Galle10] Gallè, M. "A New Tree Distance Metric for Structural Comparison of Sequences". In Apostolico, A. et al., editors, Structure Discovery in Biology, number 10231 in Dagstuhl Seminar Proc.. Leibniz-Zentrum fuer Informatik, 2010.

[ZezulaEtAl106] Zezula, P., Amato, G., Dohnal, V., and Batko, M. "Similarity Search. The Metric Space Approach". Springer, Berlin Heidelberg, Germany, 2006.



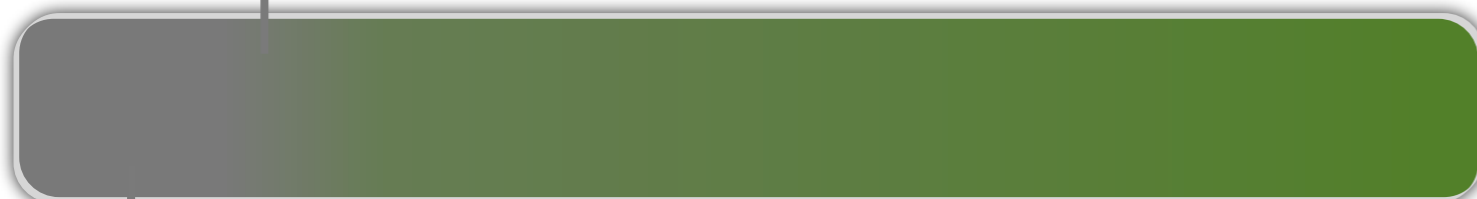
The NESTOR Prototype



The NESTOR Prototype

State of the art

**Standard
Archival
Description** [IS
AD99]



**Encoded
Archival
Description**
[Pitti01]

The NESTOR Prototype

How To Model An Archive

**Relationships with DL
Standard Technologies**

**Relationships with Archival
Standards**

**Design and Development of
the SIAR System**



The NESTOR Prototype

State of the art

Innovative

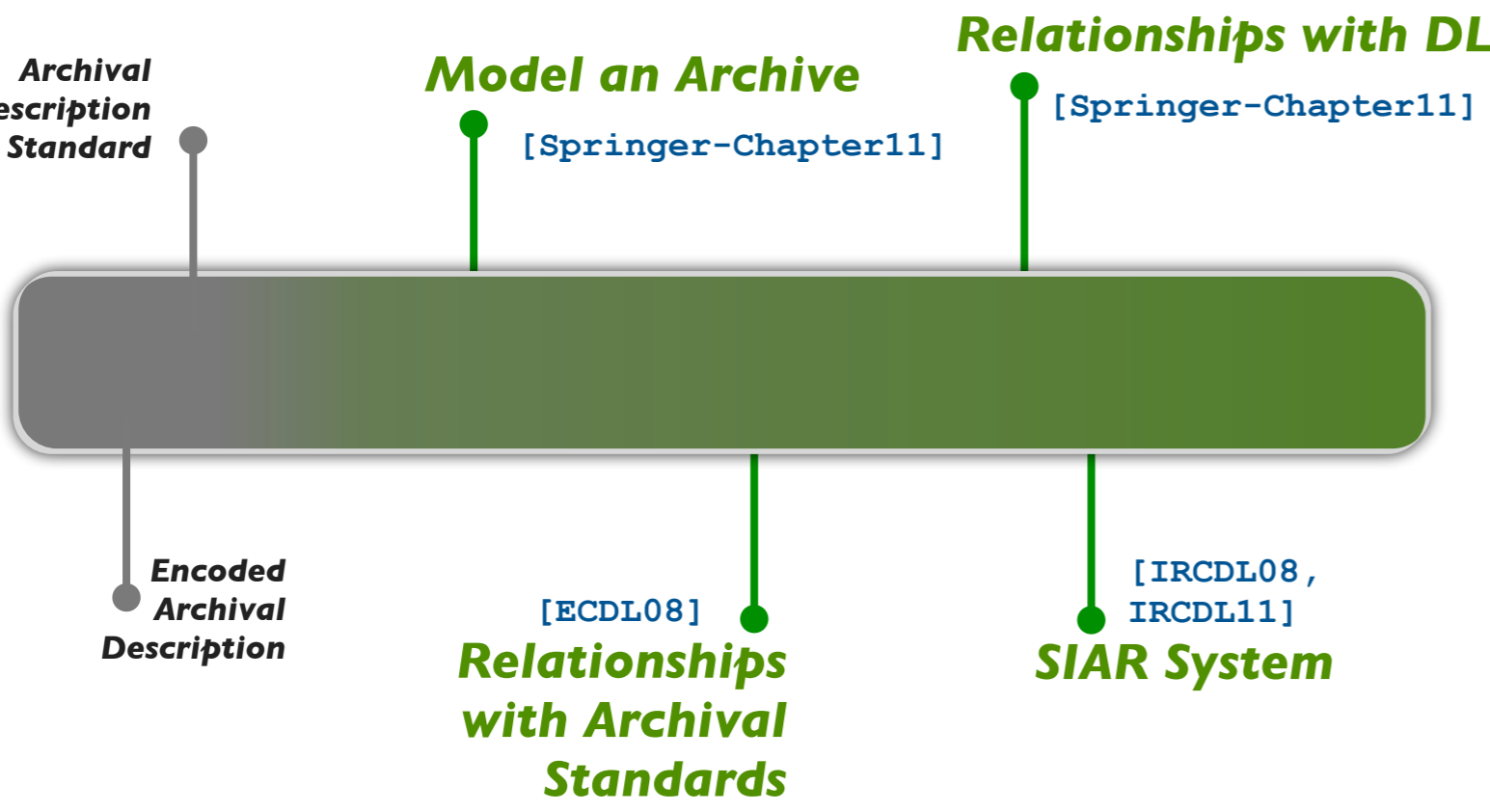
The NESTOR Prototype

How To Model An Archive

Relationships with DL Standard Technologies

Relationships with Archival Standards

Design and Development of the SIAR System

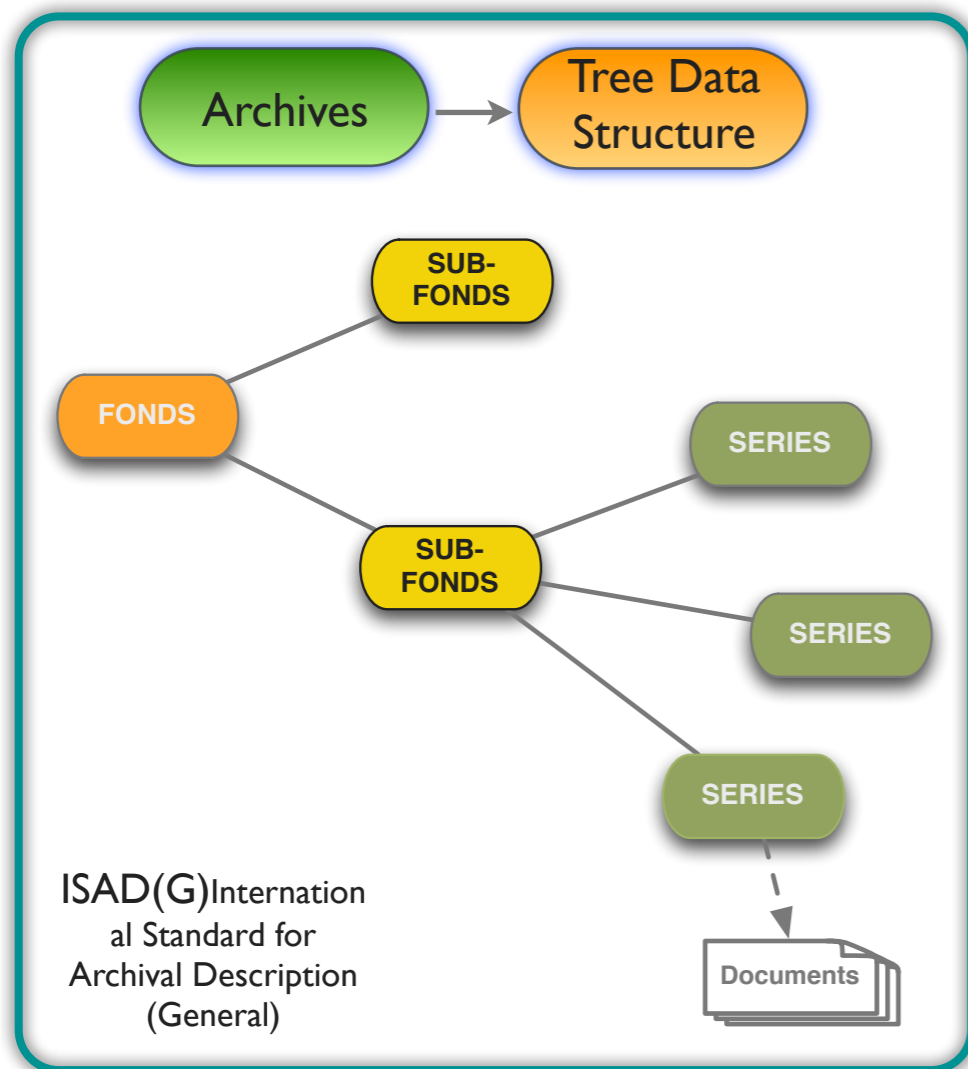


[ECDL08] Ferro, N. and Silvello, G. "A Methodology for Sharing Archival Descriptive Metadata in a Distributed Environment". In Proc.12th European Conf. on Research and Advanced Technology for DL (ECDL 2008), p.268-279. LNCS 5173, Springer, Germany, 2008.

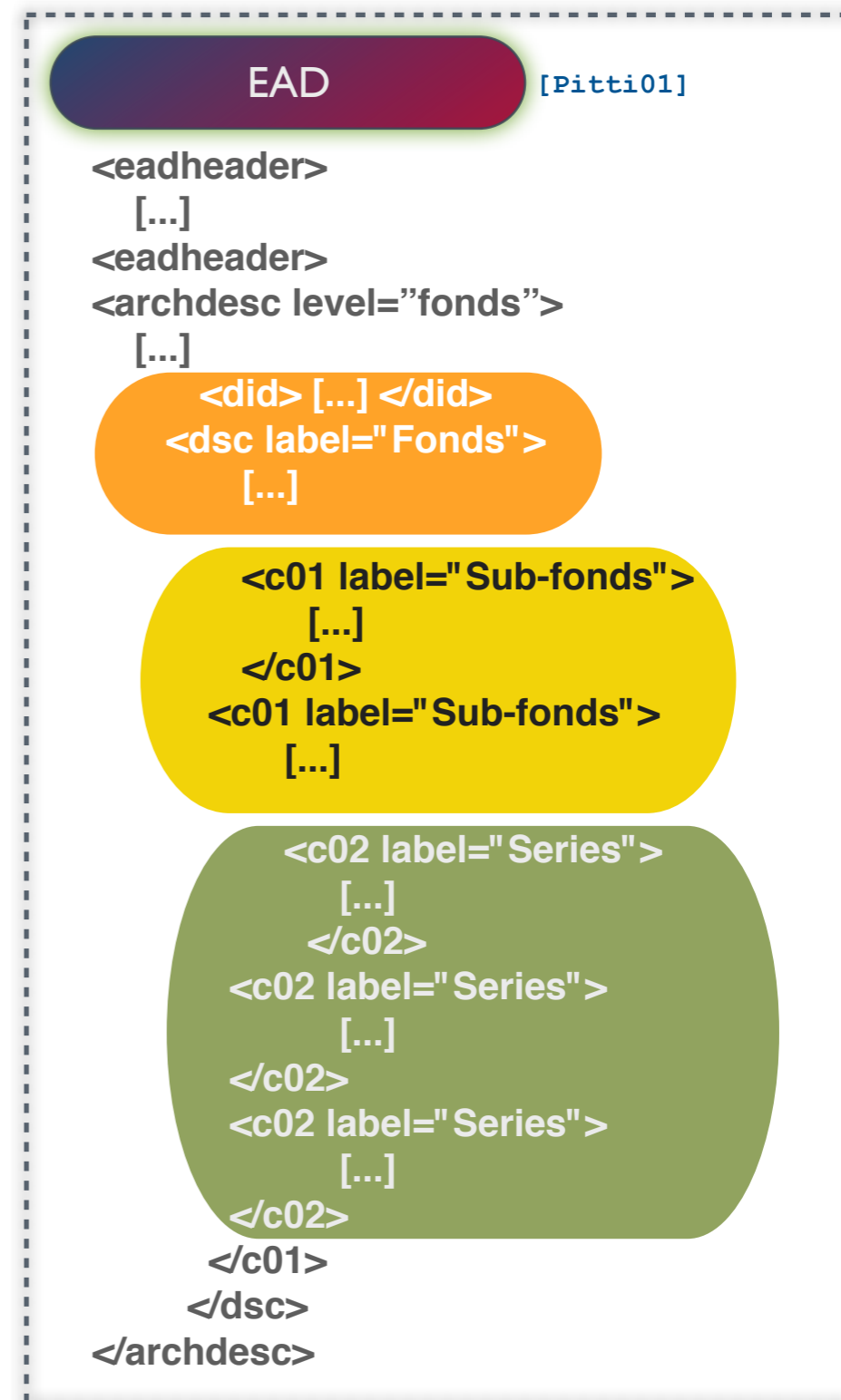
[Springer-Chapter11] Agosti, M., Ferro, N. and Silvello, G. "How to Handle Hierarchically Structured Resources Addressing Interoperability Issues in Digital Libraries". In Learning Structure and Schemas from Documents, Studies in Computational Intelligence. M. Biba and Khafa, F. eds., Springer-Verlag, Germany 2011. In print.

[IRCDL08] Ferro, N. and Silvello, G. A Distributed Digital Library System Architecture for Archive Metadata. In Agosti, M., Esposito, F., and Thanos, C., editors, Post-proceedings of the Forth Italian Research Conference on Digital Library Systems (IRCDL 2008), pages 99-104. Italy, 2008.

[IRCDL11] Agosti et al. "SIAR: A User-Centric Digital Archive System". In 7th IRCDL - Italian Research Conference on Digital Libraries, Accepted for Publication, 2011.

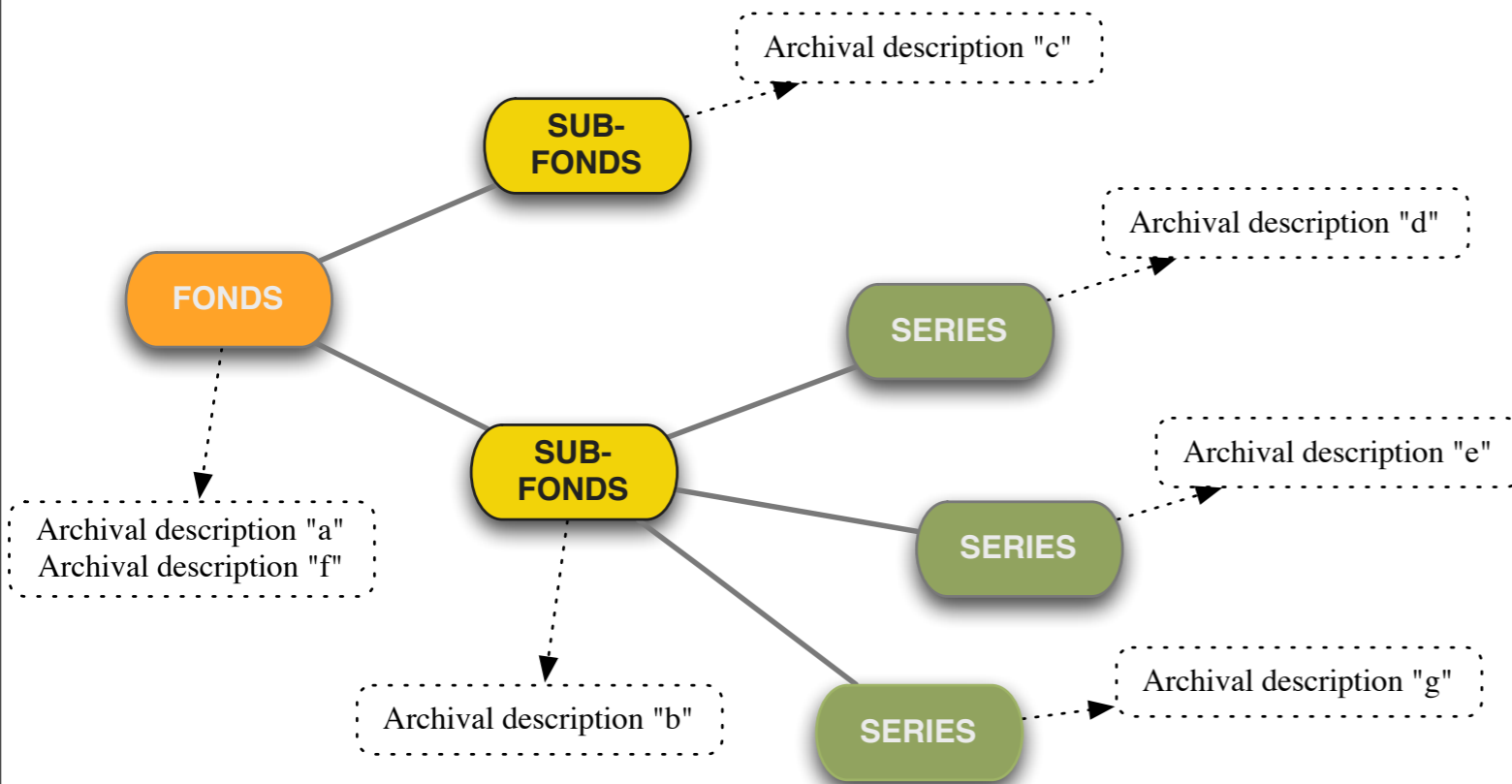


XML →

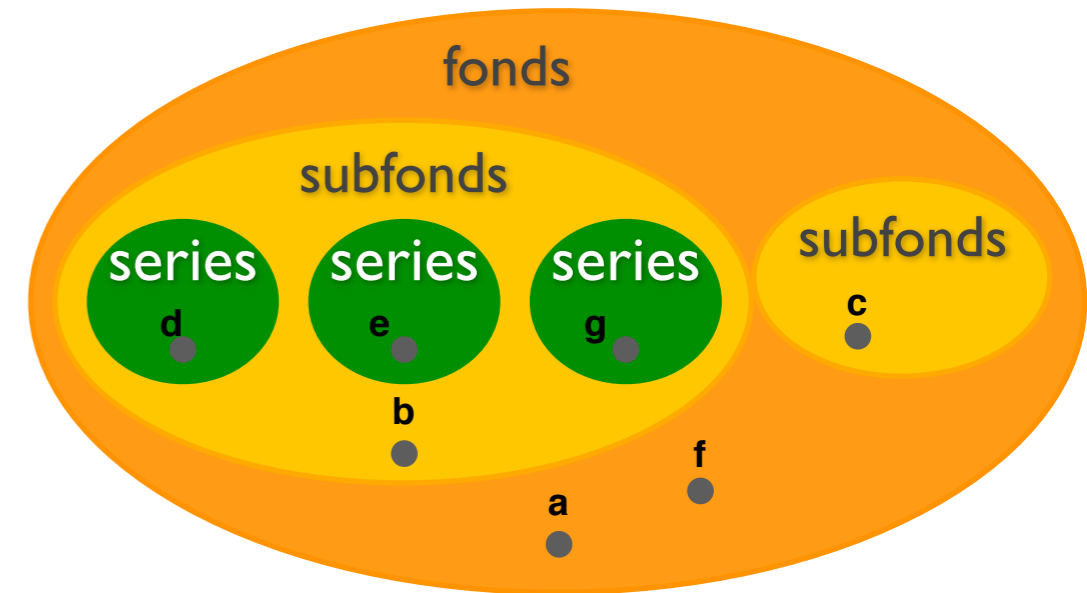


Pros and Cons [Prom02]:

- Content and Structure
- Hierarchy and Context
- Variable Granularity Access and Exchange

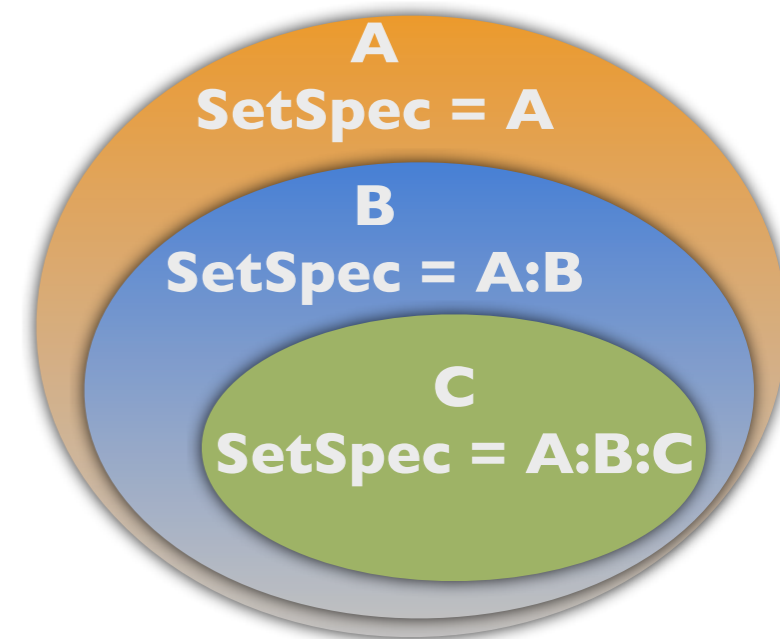
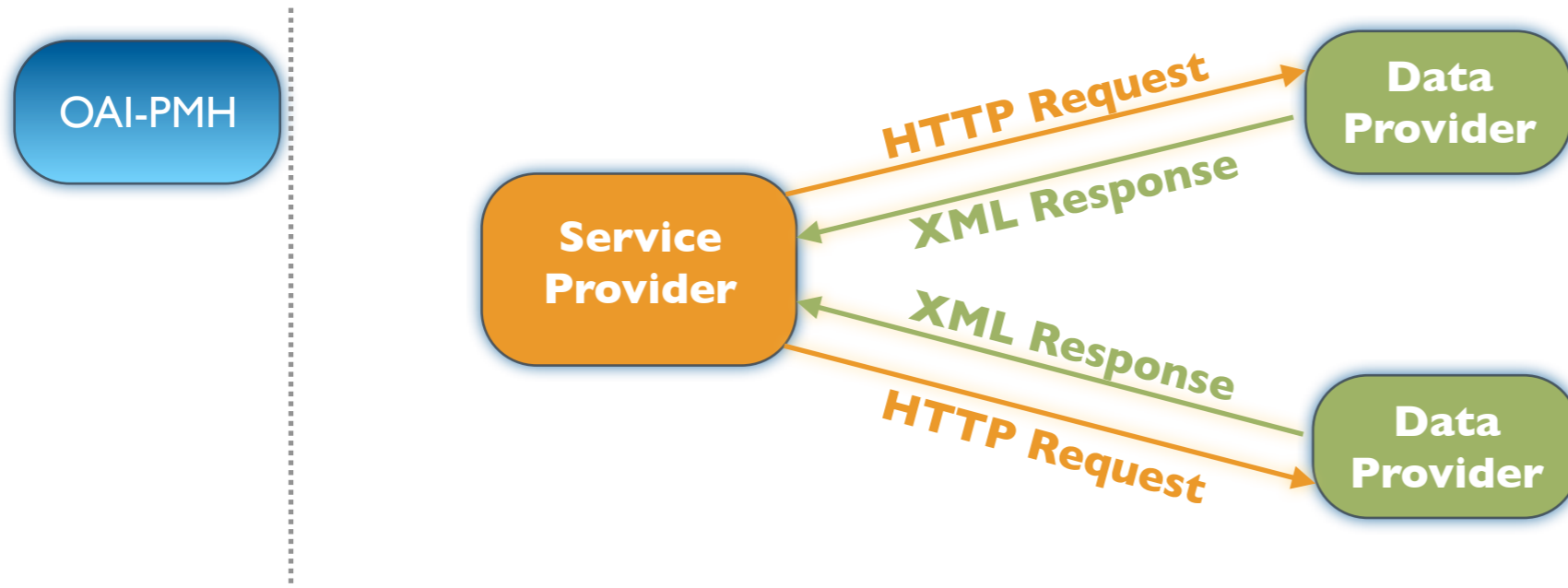


Tree Representation of an Archive



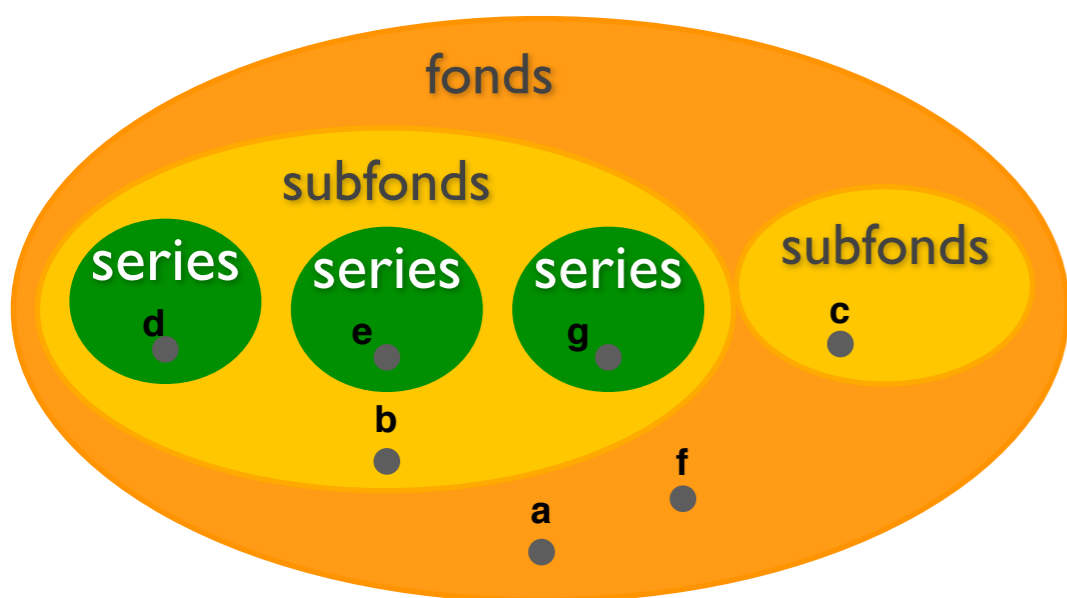
NS-M Representation of an Archive

- An archive can be modeled by means of one of the nested set models (e.g. NS-M).
- The structure is preserved by the inclusion order between the sets.
- The archival descriptions are modeled as elements belonging to the sets.
- There is a clear distinction between the structural and the content elements.



- OAI-PMH is the *standard de-facto* for metadata exchange.
- OAI-PMH enables **logical data partitioning** by defining group of records: **OAIset**.
- Harvesting from a set which has subsets will cause the repository to return **metadata** in the specified set and recursively from all its subsets [VandeSompe103, Prom03].
- **Dublin Core** is the minimum requirement of OAI-PMH.

- Digital Library technologies such as OAI-PMH and Dublin Core can be used in conjunction with the NESTOR Framework.



The metadata format is not defined by the NESTOR Framework. We can use different formats within the same set.

```
<setspec>0001</setspec>
<setname>fonds</setname>
```

```
<setspec>0001:0001</setspec>
<setname>subfonds</setname>
```

```
<setspec>0001:0002</setspec>
<setname>subfonds</setname>
```

```
<setspec>0001:0001:0001</setspec>
<setname>series</setname>
```

```
<setspec>0001:0001:0002</setspec>
<setname>series</setname>
```

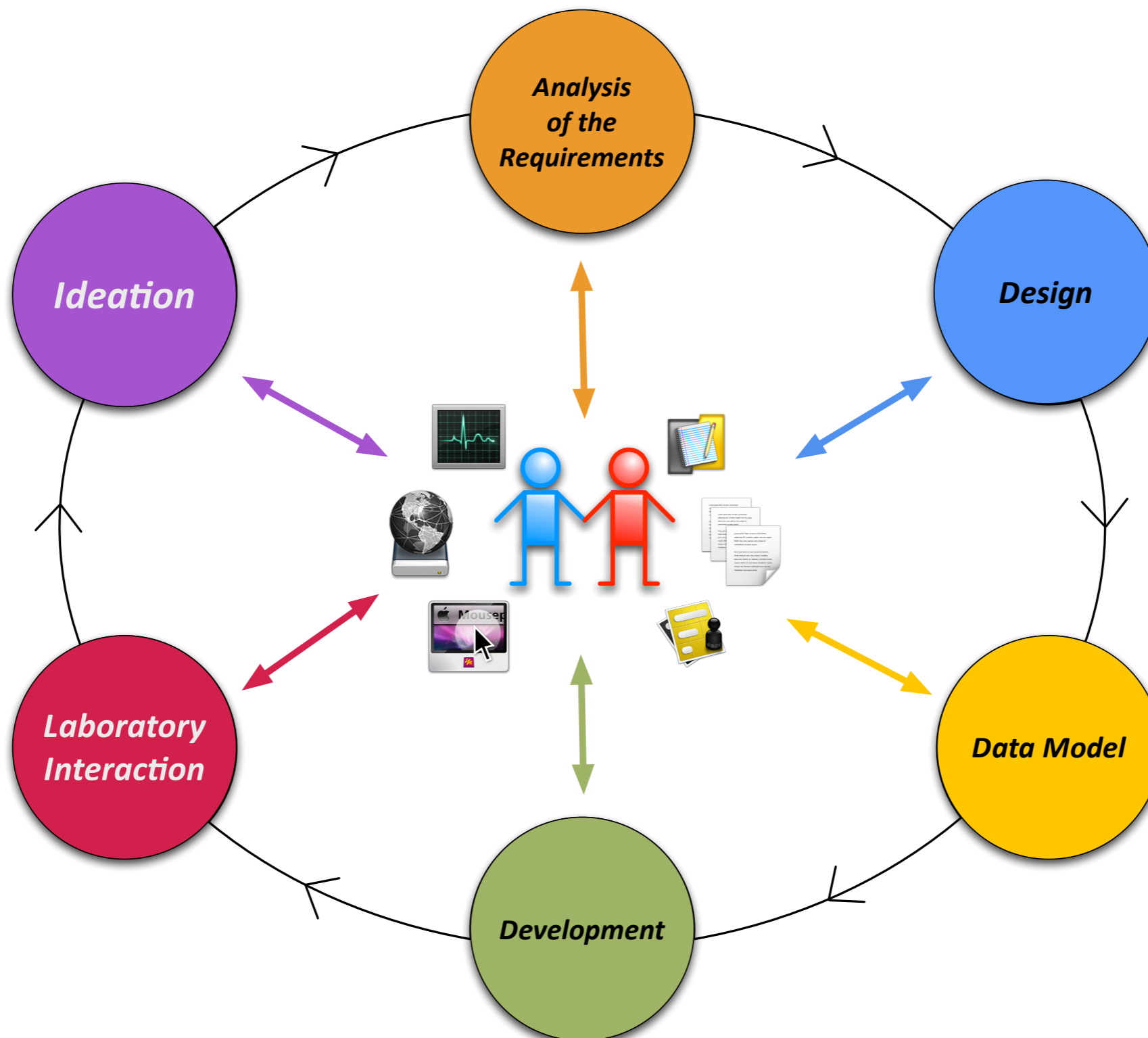
```
<setspec>0001:0001:0003</setspec>
<setname>series</setname>
```

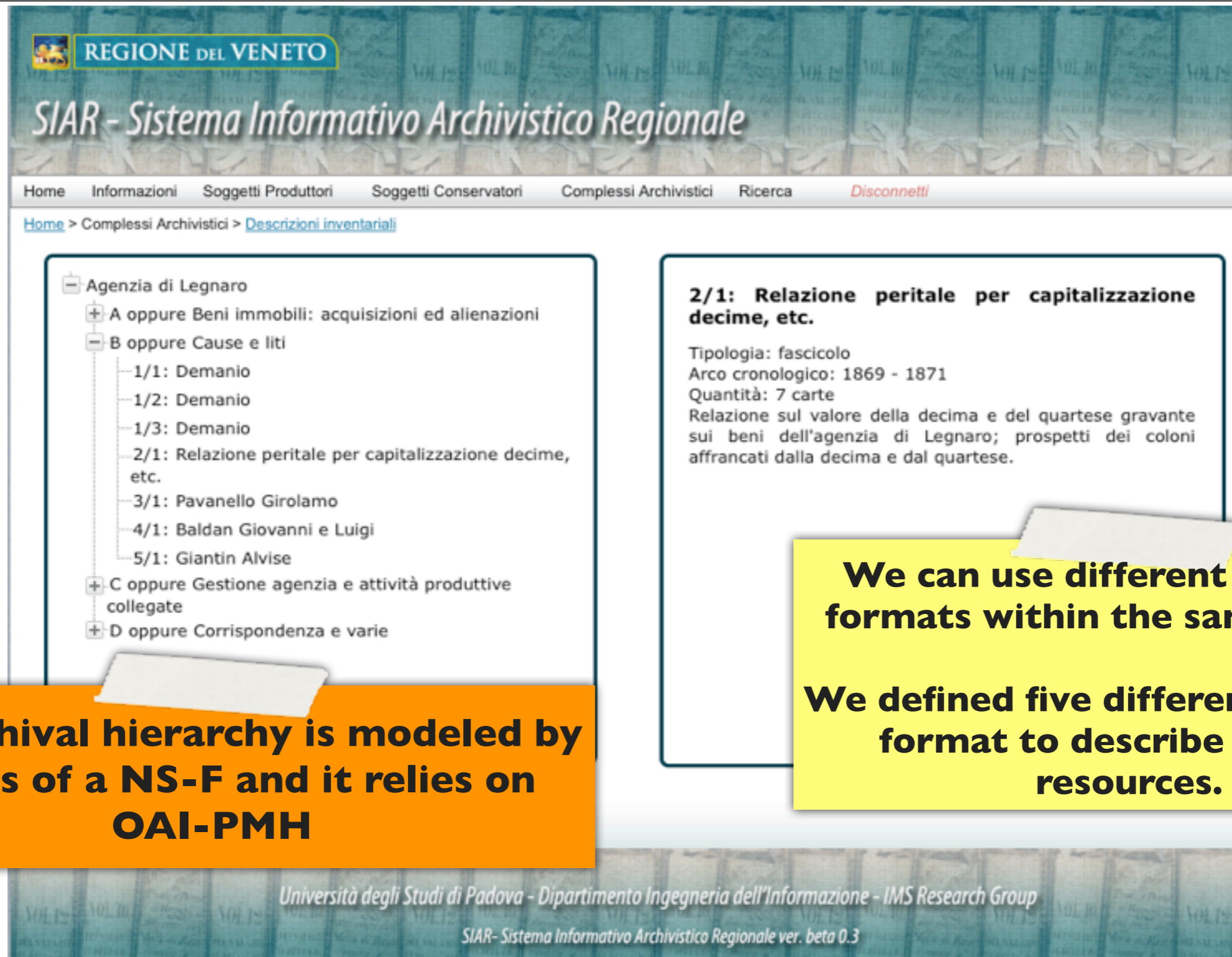
```
<record><header>
<identifier>a< /identifier>
<datestamp>2010-11-29</
datestamp><setSpec>0001</
setSpec></
header><metadata>[... ]
</metadata></record>
```

OAI Sets

OAI Records







The screenshot shows the SIAR web application interface. At the top, there is a navigation bar with links: Home, Informazioni, Soggetti Produttori, Soggetti Conservatori, Complessi Archivistici, Ricerca, and Disconnetti. Below the navigation bar, the breadcrumb trail reads: Home > Complessi Archivistici > Descrizioni inventariali. The main content area is divided into two columns. The left column displays a hierarchical tree structure for the 'Agenzia di Legnaro' with expandable/collapsible nodes. The right column displays detailed metadata for a specific record, '2/1: Relazione peritale per capitalizzazione decime, etc.', including its typology, chronological arc, quantity, and a descriptive paragraph. Two callout boxes are overlaid on the screenshot: an orange one on the left and a yellow one on the right.

REGIONE DEL VENETO

SIAR - Sistema Informativo Archivistico Regionale

Home Informazioni Soggetti Produttori Soggetti Conservatori Complessi Archivistici Ricerca *Disconnetti*

Home > Complessi Archivistici > [Descrizioni inventariali](#)

- Agenzia di Legnaro
 - + A oppure Beni immobili: acquisizioni ed alienazioni
 - B oppure Cause e liti
 - 1/1: Demanio
 - 1/2: Demanio
 - 1/3: Demanio
 - 2/1: Relazione peritale per capitalizzazione decime, etc.
 - 3/1: Pavanello Girolamo
 - 4/1: Baldan Giovanni e Luigi
 - 5/1: Giantin Alvisè
 - + C oppure Gestione agenzia e attività produttive collegate
 - + D oppure Corrispondenza e varie

2/1: Relazione peritale per capitalizzazione decime, etc.

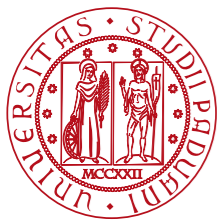
Tipologia: fascicolo
Arco cronologico: 1869 - 1871
Quantità: 7 carte
Relazione sul valore della decima e del quartese gravante sui beni dell'agenzia di Legnaro; prospetti dei coloni affrancati dalla decima e dal quartese.

The archival hierarchy is modeled by means of a NS-F and it relies on OAI-PMH

We can use different metadata formats within the same system.

We defined five different metadata format to describe archival resources.

Università degli Studi di Padova - Dipartimento Ingegneria dell'Informazione - IMS Research Group
SIAR- Sistema Informativo Archivistico Regionale ver. beta 0.3



Final Remarks and Future Work

- We addressed the research question by defining two independent **set data models** to model a hierarchy.
- We proposed an **innovative framework** (NESTOR) which re-thinks the way in which archival descriptions are approached.
- We developed a prototype of a **digital archive system** providing an **actual integration** with DL standard technologies.
- **Future work:** Define an algebra to manipulate and query the data represented by the set data models.



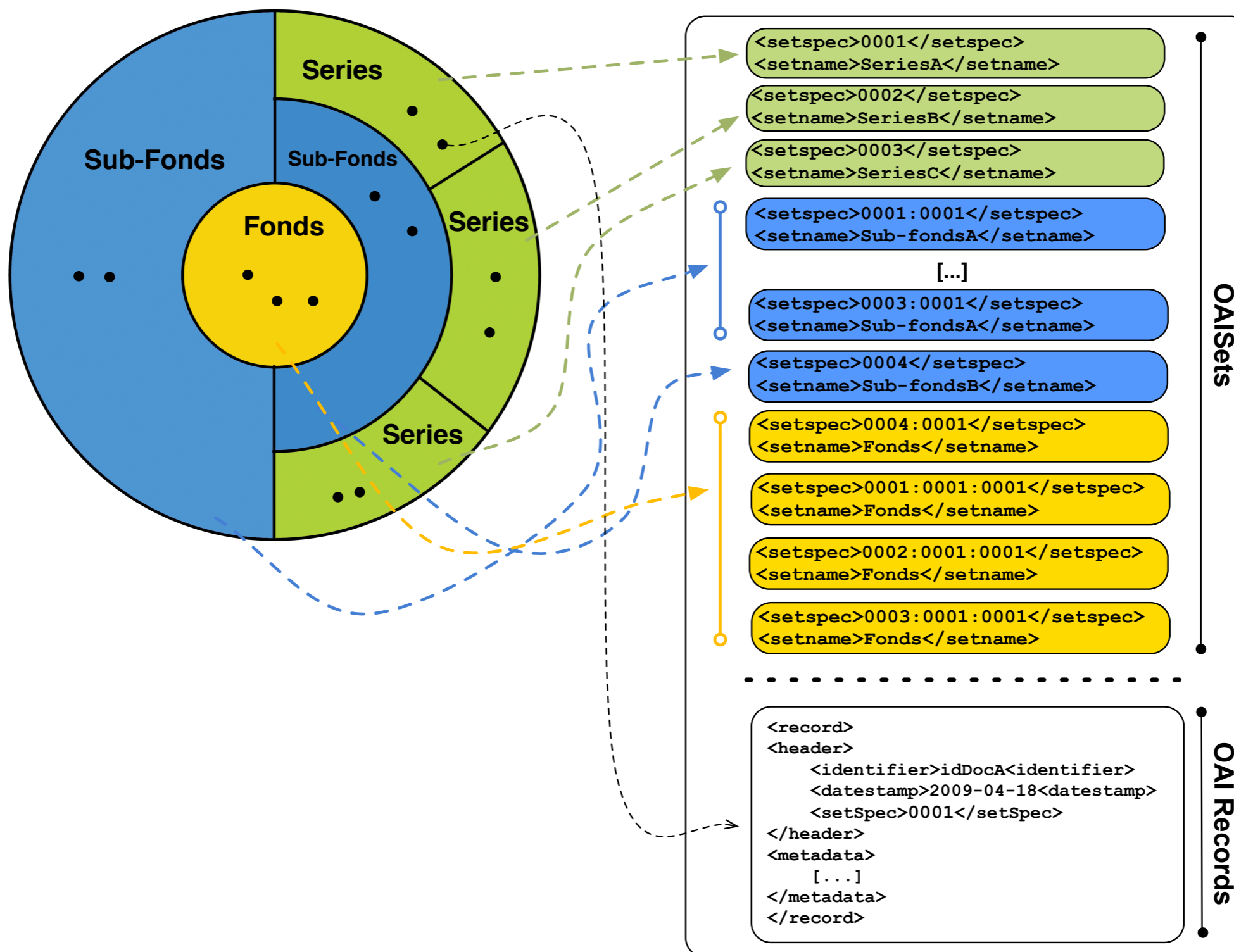
Thank You for Your Attention

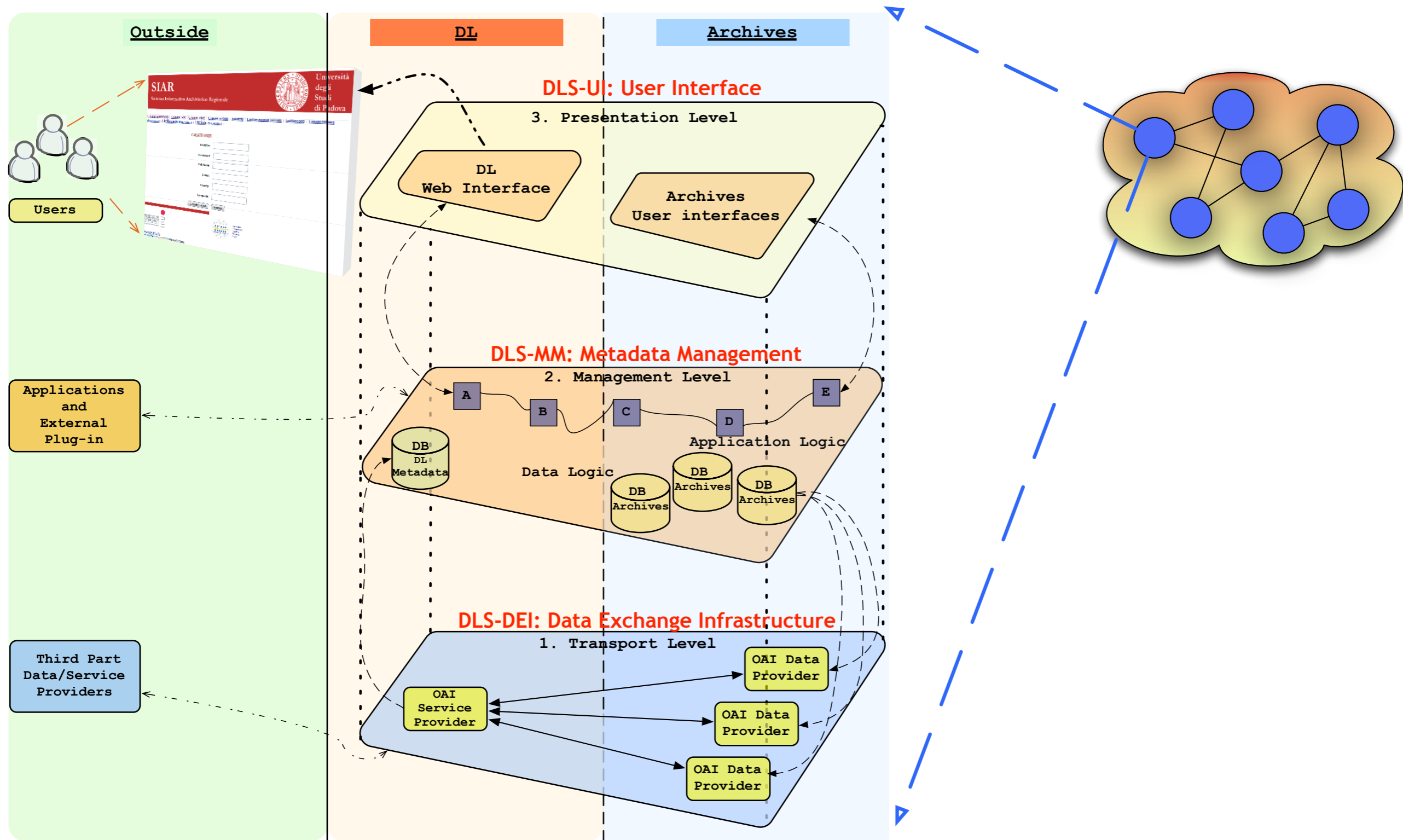


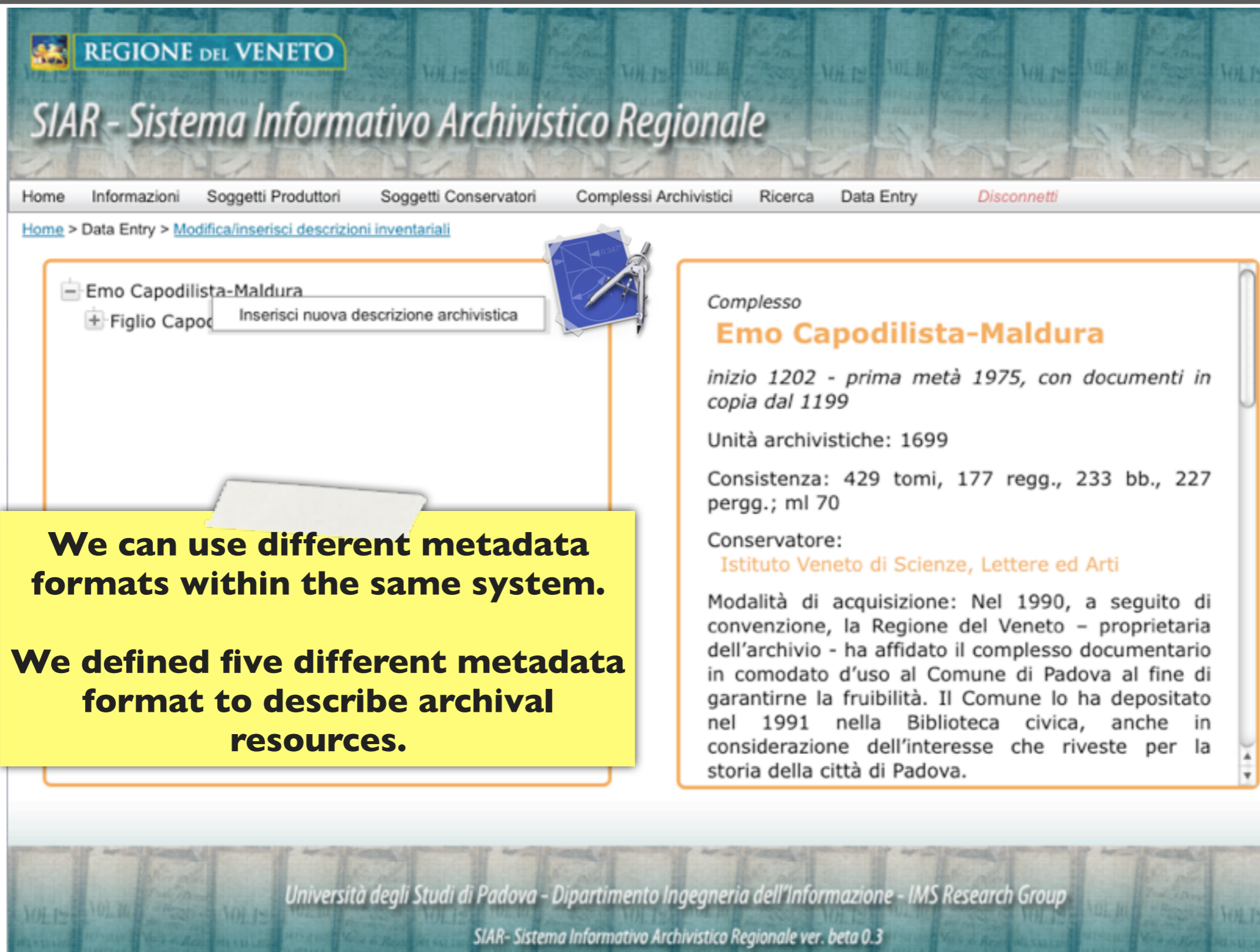
Backup slides



**Backup
Slides**







The screenshot shows the SIAR web application interface. At the top, there is a navigation menu with links: Home, Informazioni, Soggetti Produttori, Soggetti Conservatori, Complessi Archivistici, Ricerca, Data Entry, and Disconnetti. Below the menu, the breadcrumb trail reads: Home > Data Entry > Modifica/inserisci descrizioni inventariali. The main content area is divided into two sections. On the left, there is a tree view showing a folder 'Emo Capodilista-Maldura' with a sub-item 'Figlio Capoc' and a button 'Inserisci nuova descrizione archivistica'. On the right, there is a detailed record for the 'Complesso Emo Capodilista-Maldura'. The record includes the following information: 'inizio 1202 - prima metà 1975, con documenti in copia dal 1199', 'Unità archivistiche: 1699', 'Consistenza: 429 tomi, 177 regg., 233 bb., 227 pergg.; ml 70', 'Conservatore: Istituto Veneto di Scienze, Lettere ed Arti', and 'Modalità di acquisizione: Nel 1990, a seguito di convenzione, la Regione del Veneto - proprietaria dell'archivio - ha affidato il complesso documentario in comodato d'uso al Comune di Padova al fine di garantirne la fruibilità. Il Comune lo ha depositato nel 1991 nella Biblioteca civica, anche in considerazione dell'interesse che riveste per la storia della città di Padova.' At the bottom of the page, there is a footer with the text: 'Università degli Studi di Padova - Dipartimento Ingegneria dell'Informazione - IMS Research Group' and 'SIAR- Sistema Informativo Archivistico Regionale ver. beta 0.3'.

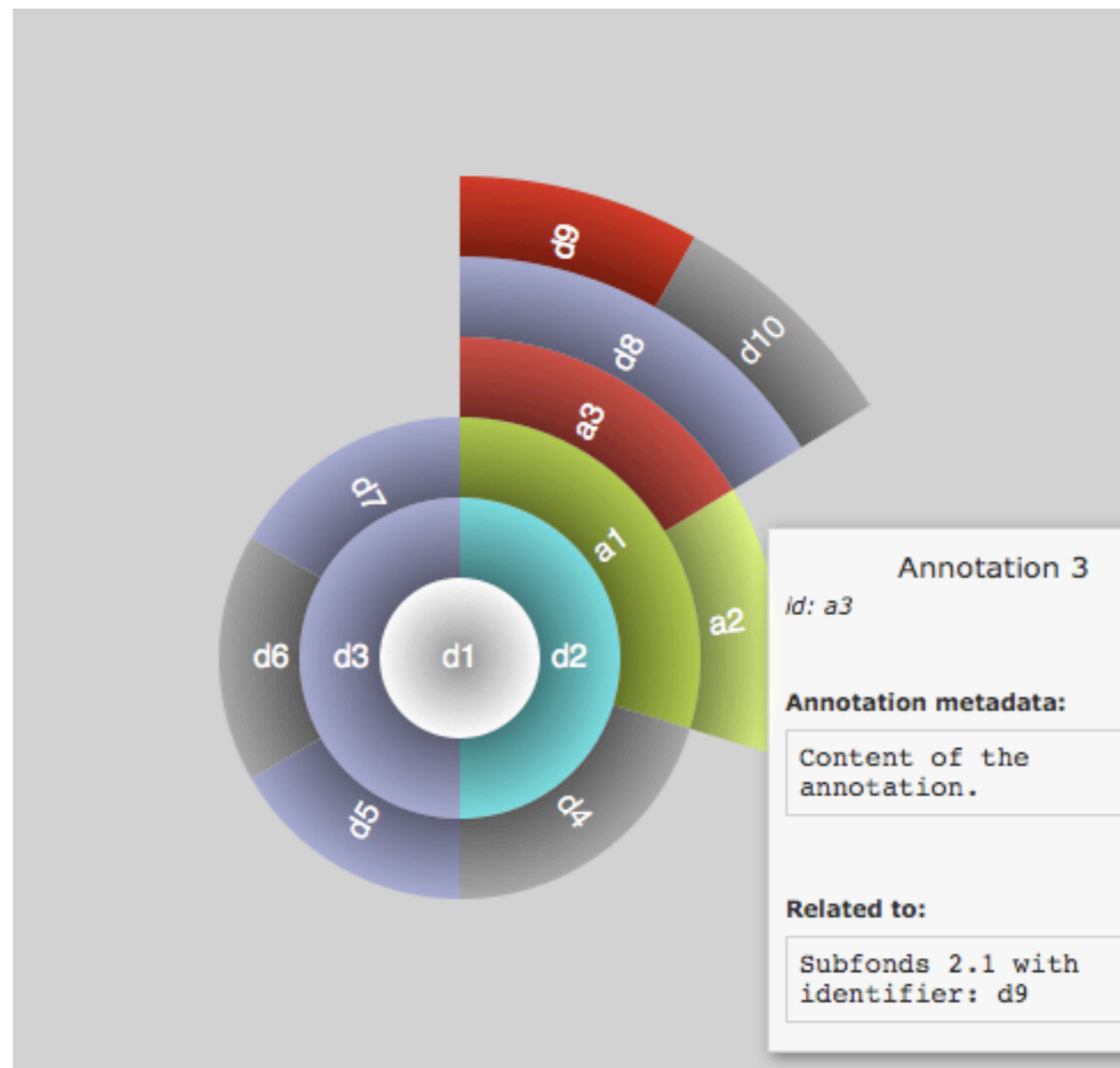
We can use different metadata formats within the same system.

We defined five different metadata format to describe archival resources.

Annotated Archives Visualization Tool

DocBall visualization of
archives and annotations.

Left click to rotate the
DocBall to the selected
circular sector and see its
details.



Sub-Fonds A

*Archival details: 1934 –
1990, with documentation
until 1994*

Archival units: 6390

Description:

Content of the
archival metadata.