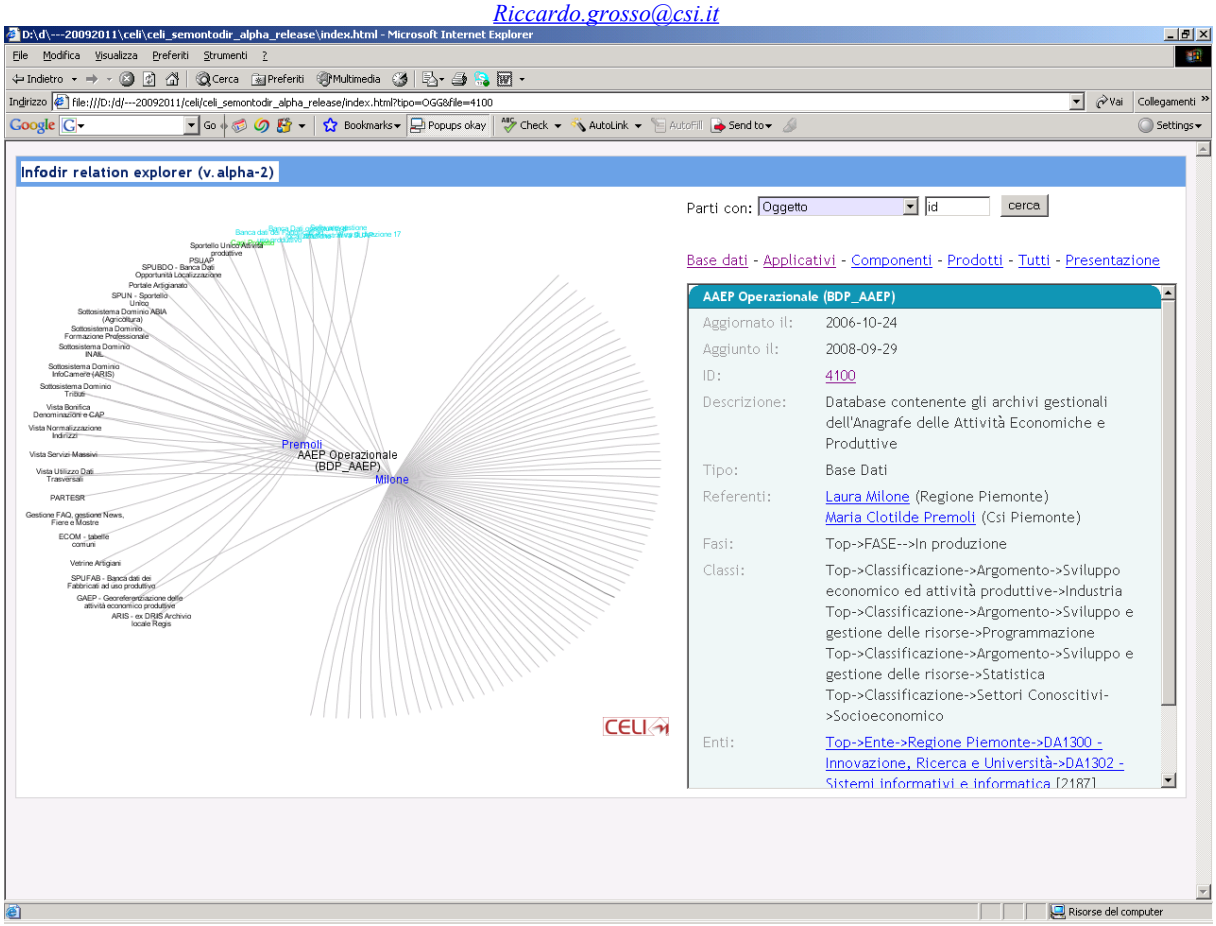
HISTORY OF CSI PIEMONTE IN METADATA CATALOGUING, *RELATION EXPLORER*, KNOWLEDGE INFERENCE AND ONTOLOGIES.

Riccardo Grosso
CSI-Piemonte
Corso Unione Sovietica 216 – 10100 - Torino - Italy
Riccardo.grosso@csi.it



ABSTRACT

CSI-Piemonte is a regional public agency with the legal status of a Consortium, organized along private lines, owned by several public stakeholders of the Piemonte area, acting in the field of information technology and telecommunication.

CSI-Piemonte, as ICT body of the Local Public Administration, manages a great deal of data, both alphanumeric and geographic, which altogether represent a wide set of detailed description of the regional data patrimony.

According to the document European Interoperability Framework (EIF), interoperability consists in the ability from part of systems ICT and supported processes to exchange data and to share information and knowledge. In particular EIF identifies three fundamental levels of interoperability to consider: organizational interoperability, semantic interoperability and technical interoperability. In this perspective the so-called semantic interoperability has a fundamental role. The definition of semantic interoperability to the inside of the document is following:

"Semantic Interoperability is concerned with ensuring that the precise meaning of exchanged information is understandable by any other application that was not initially developed for this purpose. Semantic interoperability enables systems to combine received information with other information resources and to process it in a meaningful manner."

According to the perspective of the semantic interoperability, it's not to characterize reusable data formats, but to comprise, to make exploitation, to connect between the various shadings of meant that these assume in the practical one of the administrative procedures, to the aim to render the effective composition and unification of various services possible. The technology key in order to face this problem is today that one of the so-called ontologies, data structures that they describe the meant one of the terms uses you from a computer science application or an organization so that the computers (but also the men) can reason on the relations between such terms, characterizing similarity and differences of meant.

Data and service requests are currently generated and managed in a distributed fashion. Furthermore, different actors (e.g., service providers, product sellers, governmental organizations) need to exchange data in a wealth of different formats. To allow an effective information exchange, systems need to support interoperability, thus enabling the sharing of information and knowledge. To this end, a general-purpose exchange data model needs to be defined.

The following described experiences propose a conceptual box contaning:
- in the lower parts, metadata and data of portal objects
- in the upper parts, conceptual data models ("light" ontologies)

Semantics criteria extracts knowledge from lower parts, light ontologies validates knowledge.

Upper intelligence increments lower intelligence, and viceversa.

This generalized methods allows knowledge reuse, experimented with Central Public Administration vs. Local Public Administration.

Any kinds of light ontologies, in coherence with lower level, can infer and reuse knowledge, for example also for other European Public Administration.

1. – Introduction

The metadata catalogue of the Piedmont Public Aadministration was born in 1999, with the name of ***Infodir*** (Information Directory), and like evolution of previous metadata experiences of cataloguing.
Such catalogue contains and classifies initially the business metadata of the decisional systems of the Piemonte Region.
Until 2007, technical metadata of the databases are logical/physical like ddl (data definition languages) and glossary description, produced or reversed with tools.

Beginning 2007 was released a new version of Information Directory that exceeds some architectural limits of the old system.
The new ***Infodir*** has main characteristics following:

- the backend data stewardship near the competence centers of materias in CSI and near the agencies
- separated and shared metadata views, business and technical
- generalized objects
- multidimensional model or facet-based (facets and focus)
- dynamic classifications, that is taxonomies, generalized, and combinable to criteria of text mining that they allow automatically to classify objects

To continuation of knowledge inference experience, conduced with University of Milan Bicocca and described in paper, we are placed some reflections on like generalizing it.
In other words, is possible to sofisticate criteria as an example (using the text mining or other methods) in order to make that

- increasing the conceptual knowledge based
- making criteria on whichever portal object

to obtain an increment of the semantics of the system?

2. – Ontologies Analysis within public administration

CSI-Piedmont, as ICT body of the Local Public Administration, manages a great deal of data, both alphanumeric and geographic, which altogether represent a wide set of detailed description of the regional data patrimony .

Currently Information Directory and SITAD are the metadata catalogues available for those who need to find the way through the huge amount of data (alphanumeric and geographic); these catalogues, available on RUPAR, allow the data searching either through the identification of the main topic or using keywords in the description of the name of the metadata.

Several analysis have been carried out from this rich layer of metadata:

- Knowledge inference through methods and tools which map the PA conceptual schemes with logic schemes of the data bases
- Use of ontologies, that is enriched conceptual schemes, which define the objects of the topic, the qualities and the inferences rules which allow to do a “deductive” reasoning.

The history of CSI's experience in metadata cataloguing

The Piedmonts PAL's metadata catalogue was born in 1999, with the name of ***Infodir*** (information directory); it was the evolution of previous experiences in metadata classification.

This catalogue started containing and classifying business metadata about Piedmont Region's decisional systems.

In 2002 the catalogue begins to grow and to develop along 3 main dimensions:

- public authorities and their registered metadata (not only Piedmont Region, but also City of Turin, at first, and then District of Turin and all the other local authorities);

- types of services and registered data bases, not only decisional systems, but also operational ones;

- metadata granularity, that is the introduction of data bases' technical metadata (tables and attributes) and application services (architectural components).

The main object of the metadata catalogue is the collection, to be read as the metadata cluster associated to it, that is composed of:

- data bases:
    - tables
        - t attributes
- application
    - component

Every object is equipped with a standard Dublin Core metadata set, where you may find the objects' descriptions, that are the base for both free searches and guided searches based on textual research criteria per similarity.

The main object, that is the collection, may be classified:

- by owning Public Administration
- by Topic
- by Cross section thematism

Classifications allow browsing, free researches allow searching.

Browsing and searching are fully independent and not combinable with each other

Searching modalities can be either top-down (from the collection to the database's attributes) or bottom-up (from the single data base's column to the collections and the classifications, that contain it).

Data Bases Technical metadata are photographs of the logical schemas of the data bases themselves

The catalog can be navigated by:

- Istitution (Organisation)
- Statistics (ISTAT classification)
- Cross section thematism
- Newness (backward from the most recent)
- Free research
- Word based research (headword vocabulary)
- Advanced research (using SQL criterions of equality or likeness)

Tables and columns have been catalogued with a bottom up approach, using reverse engineering techniques.

Since logical likeness mechanisms were not sufficient by themselves to make the metadata manageable, it became necessary to build the conceptual level of the catalog and to see the information representation just as users do, regardless of the physical data process.

Supertypes (entities) were built:

- with regard to the business field
- borrowed from available generalisations for central Public Administration

Hence, 261 supertypes were built, mainly concerning:

- the "business" thematism (36)
- the "subject" hierarchy (42)
- the "good" hierarchy (30)
- the "document" hierarchy (9)
- the "location" hierarchy (12)
- the "location" hierarchy enriched by CSI territorial department's contribution.

Such supertypes have been connected to 25.515 columns.

A new version of Information Directory was released in the beginning of 2007, which overcomes some architectural limits of the previous infodir of 1999.

The new infodir counts the following among its main characteristics:

- data stewardship in the subject competence centres of CSI and other bodies
- separated and shared views of metadata, both technical and of business
- generalised objects
- dimensional or facet based model
- dynamic classifications, or taxonomies, generalised and associable with text mining criterions that allow to classify the objects as they are in the repository.
- intersectable research criterions

This is the URL to reference ***infodir*** for Piemonte Region Public administration:

http://www.sistemapiemonte.it/mrspin/searchidir?type=search&term_query=&xsl=areesp3&isoptimized=on&qu_ruoliPubblici_idr=4&public=true&qu_type=obj

There are also specific links for other 2 LPA, Municipality of Turin

http://www.sistemapiemonte.it/mrspin/searchidir?type=search&term_query=&xsl=areesp3&isoptimized=on&qu_ruoliPubblici_idr=5&public=true&qu_type=obj

and Province of Turin:

http://www.sistemapiemonte.it/mrspin/searchidir?type=search&term_query=&xsl=areesp3&isoptimized=on&qu_ruoliPubblici_idr=6&public=true&qu_type=obj

Finally, its' possible to access by "CSI Piemonte view" to all LPA that have metadata catalogued on ***infodir***:

http://www.sistemapiemonte.it/mrspin/searchidir?type=search&term_query=&xsl=areesp3&isoptimized=on&qu_ruoliPubblici_idr=8&public=true&qu_type=obj

The LPA (Logical Public Administration) repository of conceptual data schemas.

In 2004, in order to increase the value of the registered metadata, a method and a tool have been experimented alongside ***Infodir***, which would have allowed the fulfilment of the following main goals:

- building an embryo of dynamic taxonomy, with "like" criteria, in order to classify the metadata for similarity with the names of the taxonomies' elements
- allowing a reciprocal exchange of inference between the used taxonomies, which are actually made with the conceptual patterns of the central Public Administration, and the constraints that are inside the structures of the registered logical data bases

In order to do this have some taxonomies been implemented, meant as entities' hierarchies and referred to the principal entities of the central PA:

- subject:
  - natural person
    - worker
      - self-employed
      - subordinate employed
        - public
      - contractor
  - juridical person (companies)

- item:
  - good
  - document

- geography:
  - place
  - territory
  - city planning

Each level of the single taxonomies has been associated with a rule of similarity, extracted from the descriptive technical metadata of the data bases' components (tables, fields).

Then also the relations between the taxonomies/hierarchies have been used, for example:

- "citizen pays tax" ("citizen" is an element of the hierarchy "natural person", "tax" is an element of the hierarchy "good")
- in order to deduce, with an top-down strategy, relations between the registered objects.

Mutually, the logical-physical objects registered in the data bases, having some reciprocal constraints between themselves, can offer bottom-up inferences, and therefore relations, between the elements of the taxonomies/hierarchies.

This taxonomic-ontologic inference's technique, used on ***Infodir***, allows to verify in which data bases each single concept of the PA is physically represented and how it is correlated, or may be correlated, both top-down and bottom-up.

In 2004, in order to improved the registered metadata patrimonium, we have experimented several methodological and design ways that permitted the development of a database Repositories creation tool.

First of all it has been studied a repository of database of Central Public Administration (CPA, developed some years ago), in order to build a specific one for the Local Administration, exploiting the similarities between the two structures.

We analysed the existing methodology to develop the tool in a first version and then we implemented it following some heuristics. The achievement of such product allows the automation of an intellectual manual work, reducing the computational time.

We can define a repository as a collection of conceptual schemas, collected by the primitives of integration and abstraction that produce in output a pyramid of schemas of the company knowledge. The conceptual schemas use a representative standard founded on Entity Relationship model that allow to show the existing relations among the objects of the system, representing the universe using classes of objects supplied with relationships and attributes.

In a period in which more and more quantity of data are manipulated by companies, a correct and functional organization of organizational system is fundamental for their efficiency; for this reason, a repository is the ideal tool to have a wide vision on the resources and to analyse the relationships between them all.
The growing of technological level in the last years introduced many governments, even the Italian one, to take care of the informatization of administrations in order to increase the quality of services for the citizens and for the businesses.
Ten years ago the first action took place for the building of the Public Administration repository; the aim was the analysis of departments' databases of Central Public Administration to create a conceptual pyramid that included the various specific knowledge. A high level of resources were necessaries to complete the CPA repository.
Nowadays, to consent to build an apparatus that allows the collaboration between the Central and Local Administration, we tried to create a repository of LPA reusing the resources of the previous experience in the CPA activity.

In the new methodology we used some heuristics to strongly reduce the computational time ad decrease the number of resources used in the job.
A first simplification is the reuse of the concepts present in the CPA conceptual pyramid to exploit the same knowledge for the reconceptualization of logical schemas of LPA, saving a lot of time compared with the work of ten years ago.

As in the case of CPA, the methodology is composed of 2 main phases:
Reconceptualization of logical schemas of databases. The knowledge supplied from the manager of Piedmont Local Public Administration is in the form of logical schemas and it cannot be used to represent concepts; it is necessary a reverse engineering operation following a specific methodology.
Integration/Abstraction. Operation of union and simplification of conceptual schemas to represent knowledge to a lower detail level.

The reconceptual phase follows a methodology composed of five elementary steps that produce pieces of knowledge that enrich incrementally the entire conceptual schemas.
We created a series of basic functions, that implement the steps in the methodology, and some superfunctions that drive them in a correct sequence:

- add entity
- add generalization
- add relationship
- add attribute
- infer constraints

To realize in the shortest time as possible a stable and complete version of the tool we decided to use some heuristics compared with the original methodology on LPA data. These simplifications concern the algorithms of each function. In this way we did not modify the structure of the methodology but only some aspects of them that could be easily modified in the future rewriting the elementary functions.

A first choice that diverges from the LPA methodology is the non-creation of a unique conceptual schema growing step by step.

This choice was made to avoid a complex building of a model that represents a conceptual schema with entity, relationships and constraints, to facilitate the firsts developmental phases.
We preferred to create simple textual output, specific for every function, that contains the information collected by its execution. The whole of these outputs, produced in the reconceptualization phase, create a set of independent information that can be singularly analysed.

The limitations we introduced have restricted the quality of the final product compared with the original LPA methodology, but we can declare that our choice guarantees a first rapid development of a stable tool.
However the presence of an expert human being is essential to verify the final quality of the results produced by the tool because the use of automations and heuristics can generate sporadic errors.

The creation of abstract schemas differs from the original LPA methodology because it was too complex to implement it in a short range of time, and this diverges from the aim of the activity.
The abstract schemas are created in gradual manner and they start from the lower part of the pyramid to reach the top following a structural hierarchy used as guide.
For this job the taxonomies that generalize concepts of LPA linked to databases are helpful; this taxonomies have the advantage to represent a specific LPA repository and they can be used as a guide to integrate and to abstract conceptual schemas.

The working contract with the realizer of tools was based on a collaboration form that facilitated a remote worker; this was caused by the distance of the author from the headquarters in Turin.
In the weekly meeting we analysed the author's work, and planned the future activity; this days produced a series of important jobs, in which the practical and theoretical knowledge of an expert person and of an apprentice were joined to produce new solutions.
As a consequence of the remote working, we have decided to optimise the implementation following the evolutive developmental methodology to facilitate the job by time and places.
In the first meeting, we discussed the choice of Editor/Compiler tool and of the DBMS; we decided to use Microsoft Visual Basic 6 and Microsoft Access because of the simplicity and good knowledge of these applications.

Then, we collected the knowledge of LPA. Almost the whole data were exported from the central system (***InfoDir***) to the database managed by DBMS Access to be easily manipulated with query SQL to supply specific request.

In the developmental phase we produced the five steps for the reconceptualization of a logical schema of a LPA database and the unique step for integration-abstraction.
At the end of this phase of development of the elementary functions, we could design and implement the superfunctions that drive the basic functions in a correct logical sequence.

The tool is subdivided in three macro areas, corresponding to the user functions:

- Reconceptualization of a database
- Integration-abstraction of schemas
- Creation of a repository

As it is easy to understand, the areas grow linearly by complexity and recall concepts of the previous areas.

As far as the interest shown in the project was very high, some upgrading operations were planned with the aim to increase the quality of the final product.
We can divide the improvements in two directions: the correctness of the contents of a conceptual schema and a better representation.

To conclude, the activity have satisfied the requests, it has supplied an efficient tool with a good degree of efficacy in regards of the contest, and it has settled important bases for future researches.

3. – Conclusion

In order to describe possible extends and reuse of our experience in other Projects, we consider this aspects:

- our metadata and schema repository, like a blackbox, consist of
    - "light ontologies" in upper part of blackbox (conceptual schemas with glossaries) valid on CPA
    - semantic search criteria on lower part (portal objects with descriptive metadata, not only database tables and field) valid on LPA
    - "light ontologies" in middle part, obtained with methodology and tool described, like a "marriage" from CPA concepts and LPA concepts inferred

So that other Italian region (other italian LPA) can reuse methodology and tool for analogue "marriage" with Italian CPA concepts, also other European CPA with corresponding LPA can reuse the solution.

It is sufficient to "change the input" to the blackbox:
- other CPA light ontologies
- other LPA portal objects to infer

Methods and tools described, can be reused theorically for any domain, for example:
- marketing light ontologies in blackbox upper part
- objects of marketing portal

References

Follows some references about presentation and publication of our works with prof. Carlo Batini of Milan Bicocca University:

Kumar, S.: Data governance: An approach to effective data management. White paper, Satyam Computer Services, Ltd. (2008)
[2] Hauch, R., Miller, A., Cardwell, R.: Information intelligence: metadata for information discovery, access, and integration. In: SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, New York, NY, USA, ACM (2005) 793–798
[3] Mei, J., Xie, G.T., Zhang, L., Liu, S., Schloss, R.J., Pan, Y., Ni, Y.: Umrr: Towards an enterprise-wide web of models. In Bizer, C., Joshi, A., eds.: International Semantic Web Conference (Posters & Demos). Volume 401 of CEUR Workshop Proceedings., CEUR-WS.org (2008)
[4] Batini, C., Battista, G.D., Santucci, G.: Structuring primitives for a dictionary of entity relationship data schemas. IEEE Trans. Softw. Eng. 19 (1993) 344–365
[5] Staab, S., Studer, R.: Handbook on Ontologies (International Handbooks on Information Systems). SpringerVerlag (2004)
[6] Hepp, M.: Possible ontologies: How reality constrains the development of relevant ontologies. IEEE Internet Computing 11 (2007) 90–96
[7] Calvanese, D., Lembo, D., Lenzerini, M., Rosati, R.: Dl-lite: Tractable description logics for ontologies. In: In Proc. of AAAI 2005. (2005) 602–607
[8] Palmonari, M., Batini, C.: Representing and integrating light-weight semantic web models in the large. In: Proceedings of the 5th International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS 2009). (2009)
[9] Calvanese, D., Lenzerini, M., Nardi, D.: Unifying

class-based representation formalisms. J. of Artificial Intelligence Research 11 (1999) 199–240
[10] Xu, Z., Cao, X., Dong, Y., Su, W.: Formal approach and automated tool for translating er schemata into owl ontologies. In Dai, H., Srikant, R., Zhang, C., eds.: PAKDD. Volume 3056 of Lecture Notes in Computer Science., Springer (2004) 464–475
[11] Artale, A., Calvanese, D., Kontchakov, R., Ryzhikov, V., Zakharyaschev, M.: Reasoning over extended er models. In: Proc. of the 26th Int. Conf. on Conceptual Modeling (ER 2007). Volume 4801 of Lecture Notes in Computer Science., Springer (2007) 277–292
[12] Batini, C., Barone, D., Garasi, M., Viscusi, G.: Design and use of er repositories: Methodologies and experiences in egovernment initiatives. In Embley, D.W., Oliv´e, A., Ram, S., eds.: ER. Volume 4215 of Lecture Notes in Computer Science., Springer (2006) 399–412
[13] Paslaru, E., Simperl, B., Tempich, C., Sure, Y.: Ontocom: A cost estimation model for ontology engineering. In: In Proceedings of the 5th International Semantic Web Conference ISWC2006. (2006)
[14] Tavana, M., Joglekar, P., Redmond, M.A.: An automated entity-relationship clustering algorithm for conceptual database design. Inf. Syst. 32 (2007) 773–792
[15] Mylopoulos, J., Borgida, A., Jarke, M., Koubarakis, M.: Telos: Representing knowledge about information systems. ACM Trans. Inf. Syst. 8 (1990) 325–362
[16] Calvanese, D., Giacomo, G.D., Lenzerini, M., Nardi, D., Rosati, R.: Description logic framework for information integration. In: KR. (1998) 2–13
[17] Noy, N.F.: Semantic integration: a survey of ontology-based approaches. SIGMOD Rec. 33 (2004) 65–70
[18] Batini, C., Ceri, S., Navathe, S.B.: Conceptual database design: an Entity-relationship approach. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA (1992)
[19] Castano, S., Antonellis, V.D., Fugini, M.G., Pernici, B.: Conceptual schema analysis: techniques and applications. ACM Trans. Database Syst. 23 (1998) 286–333
[20] Campbell, L.J., Halpin, T.A., Proper, H.A.: Conceptual schemas with abstractions making flat conceptual schemas more comprehensible. Data Knowl. Eng. 20 (1996) 39–85
[21] Dahchour, M., Pirotte, A., Zim´anyi, E.: Generic relationships in information modeling. 3730 (2005) 1–34
[22] Keet, C.M.: Enhancing comprehension of ontologies and conceptual models through abstractions. In: AI*IA '07: Proceedings of the 10th Congress of the Italian Association for Artificial Intelligence on AI*IA 2007, Berlin, Heidelberg, Springer-Verlag (2007) 813–821
[23] Wang, Z., Wang, K., Topor, R.W., Pan, J.Z.: Forgetting concepts in dl-lite. In Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M., eds.: ESWC. Volume 5021 of Lecture Notes in Computer Science., Springer (2008) 245–257
[24] Sousa, P., de Jesus, L.P., Pereira, G., e Abreu, F.B.: Clustering relations into abstract er schemas for database reverse engineering. Sci. Comput. Program. 45 (2002) 137–153
[25] Chen, P.P.S.: The entity-relationship model—toward

a unified view of data. ACM Trans. Database Syst. 1 (1976) 9–36
[26] Ghidini, C., Kump, B., Lindstaedt, S.N., Mahbub, N., Pammer, V., Rospocher, M., Serafini, L.: Moki: The enterprise modelling wiki. In Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyv¨onen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E.P.B., eds.: ESWC. Volume 5554 of Lecture Notes in **Computer Science., Springer (2009) 831–835**

http://www.via-nova-architectura.org/files/EMMSAD2005/Batini.pdf
http://books.google.it/books?id=kjt5JW6ZSdYC&pg=PA170&lpg=PA170&dq=batini+grosso&source=web&ots=Z7Vo7jAs8D&sig=70jOYba7eQ_zb-Mf-RK_P87EQB0&hl=it
http://www.inderscience.com/search/index.php?action=record&rec_id=9601&prevQuery=&ps=10&m=or
http://www.iseing.org/egov/eGOV05/Source%20Files/Papers/CameraReady-7-P.pdf
http://www.urbanontology.net/Programme.htm
http://www.iasummit.it/download/09-siias2007.ppt