MONTRA: An agile architecture for data publishing and discovery

Luís Bastião Silva, Alina Trifan, José Luís Oliveira

PII: S0169-2607(17)31272-5
DOI: 10.1016/j.cmpb.2018.03.024
Reference: COMM 4664

To appear in: *Computer Methods and Programs in Biomedicine*

Received date: 21 October 2017
Revised date: 26 February 2018
Accepted date: 27 March 2018

MONTRA: An agile architecture for data publishing and discovery

Luís Bastião Silva[1], Alina Trifan[2], José Luís Oliveira[2]

[1] BMD Software, Aveiro, Portugal,
bastiao@bmd-software.com
[2] University of Aveiro, DETI/IEETA, Portugal,
{alina.trifan, jlo}@ua.pt

Abstract

Background and Objective: Data catalogues are a common form of capturing and presenting information about a specific kind of entity (e.g. products, services, professionals, datasets, etc.). However, the construction of a web-based catalogue for a particular scenario normally implies the development of a specific and dedicated solution. In this paper, we present MONTRA, a rapid-application development framework designed to facilitate the integration and discovery of heterogeneous objects, which may be characterized by distinct data structures.
Methods: MONTRA was developed following a plugin-based architecture to allow dynamic composition of services over represented datasets. The core of MONTRAs functionalities resides in a flexible data skeleton used to characterize data entities, and from which a fully-fledged web data catalogue is automatically generated, ensuring access control and data privacy.
Results: MONTRA is being successfully used by several European projects to collect and manage biomedical databases. In this paper, we describe three of these applications scenarios.
Conclusions: This work was motivated by the plethora of geographically scattered biomedical repositories, and by the role they can play altogether for the understanding of diseases and of the real-world effectiveness of treatments. Using metadata to expose datasets' characteristics, MONTRA greatly simplifies the task of building data catalogues. The source code is publicly available at `https://github.com/bioinformatics-ua/montra`.

Keywords: biomedical databases, data catalogues, patient registries, clinical studies

1 Introduction

Data integration methodologies have become crucial in many research fields, and particularly in biomedical sciences, by enabling the discovery of knowledge that leads to new advancements in research, as well as by stimulating adoption of the necessary technical mechanisms to make such processes traceable, shareable and integrable. However, integrating biomedical data is a major challenge faced by applications that need to query across multiple autonomous data sources [1, 2]. A reliable data integration system must deliver data from a variety of sources and must cope with any limitations that the data may impose [3].

Biomedical research represents a large field that can greatly benefit from data integration tools. Clinical studies, for instance, could be conducted in a faster and more extensive way if data integration systems could provide not only the integrated data of different biomedical datasets, but also the tools needed in order to discover, compare and consult the datasets that can support the study. However, due to the numerous challenges that the integration of biomedical data imposes, data sharing is not the default, but the exception [4].

One of the most common challenges to be overcome is that biomedical data sources are often hard to locate or unavailable outside of the institution that owns them. Moreover, data privacy is an important aspect to be taken into consideration and often a limiting factor when it comes to clinical information sharing [5]. Biomedical data integration systems have to provide solutions for integrating data from multiple sources without having to first load all the data into a central warehouse, since this would not only be impracticable, but would also raise many ownership and privacy issues. Another key challenge resides in finding ways of dealing with the heterogeneity, diversity and complexity of the information found in geographically scattered databases and medical healthcare units [6]. However, the value of any kind of data is greatly enhanced when it exists in a form that allows it to be integrated with other data [7].

Current efforts go towards defining data standards and models [8, 9], common ontologies and semantics [10] that can support fluid data integration. In this paper we propose an alternative solution and we intend to tackle the biomedical data integration problem from a different perspective. The architecture we propose provides a different view of biomedical data integration, in which integration and sharing is made possible independently of the structure of the biomedical data entities and the type of data they contain.

We introduce a Rapid Application Development system [11], designated MONTRA, which is intended as a sustainable framework, capable of enabling data linkage at a level of detail not currently available in any other systems. The core of MONTRA's functionalities resides in a dynamic data skeleton used for the characterization of data entities, or in other words, metadata extraction. These metadata can be browsed, compared and queried by users of the system without having any private data exposed.

This paper is divided into 5 more sections. In the following section we present an overview of several data catalogue solutions, especially the ones being used in the biomedical area. In section 3.1 we describe several key requirements that should be fulfilled by a data integration system. MONTRA's architecture is detailed in Section 4.1, while Section 5 discusses the use of this system within three different scenarios. Finally, in the conclusions we discuss the main outcomes of this work.

2 Background

Based on large and commonly supported research infrastructures, universal computational platforms capable of providing unified solutions for multiple life science needs are emerging [12]. These solutions, complemented by open source and open access policies, have the potential to sustain the development of data integration computational systems. Integrated data fosters knowledge that can support not only advances in biomedical research, but that can also significantly improve patient care, public health and administrative efficiency [13]. Most of the challenges these systems have to overcome are data size, heterogeneity, geographical location and data privacy. In this section we will review some of the current solutions for biomedical data integration.

Cohort discovery platforms, as the name suggests, focus on the discovery of cohort data sources, usually related to a specific disease. The Global Alzheimers Association Interactive Network (GAAIN) [14], for instance, aims to accelerate the development of Alzheimers disease prevention, treatments and a cure. The platform fosters cohort discovery, collaboration and sharing. The i2b2 project [15] enables researchers to discover cohorts of patients using data from Electronic Health Record systems. In addition, it supports different types of queries of clinical data, including whether clinical concepts occurred at any point in a patients medical history, during a particular visit, or in a sequence of events. In the area of genomics, the Cohort Discovery [16] and the Genomics Cohort Catalogue [17] are web platforms that connect a wide range of research cohort data with resources and specimens available for further investigation. An extensive range of well phenotyped and catalogued population cohorts representing more than 600,000 subjects and including a number of ethnically homogeneous population sets can be reached through the ENGAGE Catalogue (European Network for Genetic and Genomic Epidemiology) [18]. A more sophisticated search and query data discovery platform is CafeVariome [19]. The platform is built as a shop window interface to support the discovery of genotype-phenotype data and to allow data access under three different models.

Another wide topic addressed by these health science catalogues is clinical studies. CLOSER Discovery [20] is a search engine that allows researchers to explore the content of eight leading UK longitudinal studies. A similar tool, the Quebec Study Catalogue [21] is a Maelstrom Research initiative aiming to document and promote the scientific usage of large scale epidemiological studies in Quebec. With a wider impact, Clinicaltrials.gov [22] is a database of privately and publicly funded clinical studies conducted around the world. The ImmPort [23] project focuses on archiving and exchanging research and clinical data for the life science researchers. At its core, ImmPort is a data warehouse containing experimental data and metadata that describe the purpose of a study and the methods of data generation. A more advanced web platform, MOLGENIS [24] was developed from molecular genetics research and has been used in several scientific areas such as biobanking, rare disease research and patient registries. It comprises a suite of web databases for genotype, phenotype and analysis pipelines, as well as a software generator to rapidly build web databases.

Among the phenotype and genotype integration platforms, the Monarch Initiative [25] provides a portal for exploration of phenotype-based similarity. For this it integrates and re-distributes cross-species gene, genotype, variant, disease, and phenotype data.

With a broader approach, the Catalogue of Activities in eHealth [26] was designed to identify and aggregate global resources about clinical and genomic data. The catalogue enables both researchers and clinicians to find appropriate resources to meet the needs of their data sharing projects. Improving access, facilitating the secondary use of health data and providing technical and governance solutions are among the aims of the EMIF project [27]. To this end, a common information framework (EMIF-Platform) links up and facilitates access to diverse medical and research data sources. Similar to the efforts envisioned by the EMIF Platform, ELIXIR [28] is an intergovernmental organization that brings together life science resources from across Europe, such as databases, software tools and training materials. The organization's goal is to coordinate them into a single infrastructure that could enable finding and sharing data and expertise. Another collaborative scientific network is the ENCePP (European Network of Centres for Pharmacoepidemiology and Pharmacovigilance), whose goal is to strengthen the post-authorisation monitoring of medicinal products, whilst bringing together relevant research centers, healthcare databases, electronic registries and existing networks across Europe [29]. Throughout the US, two of the best-known initiatives dedicated to the integration of health databases are Bridge-To-Data [30] and the Healthcare Cost and Utilization Project (HCUP) [31]. Bridge-To-Data offers services that allow users to identify key features and compare database profiles, while it also serves as an educational tool for public health research. HCUP includes the largest collection of longitudinal hospital care data in the United States. Researchers and policy-makers use HCUP data to identify, track, analyze and compare hospital statistics at the national, regional and State levels. While being able to collect virtually any type of data. REDCap (Research Electronic Data Capture) [32, 33] is specifically geared to support online or offline data capture for research studies and operations. It allows researchers to define project-specific data capture and launch protocol data collection.

As previously discussed, different approaches for data capture, integration and analysis of biomedical data have been designed. Some of them focus on a specific disease, health topic, or geographical location of the data sources, while others offer a wider access to distinct resources. To facilitate the construction of these repositories, we propose MONTRA, an out-of-the-box architecture for designing data integration platforms, with emphasis on biomedical data. Such platforms can cover a large spectrum of biomedical data sources, that can eventually be organized by research topic or disease. A system based on MONTRA is able to centralize heterogeneous biomedical data sources under the same web interface, thus making them accessible from a unique entry point. The main advantages of MONTRA resides in the ability of building such web platforms almost on the fly, as well as on the flexibility that this architecture offers and that we will discuss in more detail in the next sections.

3 Methods

3.1 System Requirements

From the solutions and challenges of the systems presented in the previous section we have found a set of requirements that should be addressed within the architecture that we present.

- *Flexibility* - The system should be flexible enough to integrate information coming from different sources. This information might be stored on diverse physical storage devices, obey distinct data models and concern more than just one specific research topic. Addressing the variety of information that should be managed by such a system is a critical aspect that has already been identified in the literature [10].
- *Data Privacy* - Despite enabling collaborative research and clinical data reuse, no private data can be exposed. The system must be able to perform biomedical data integration while taking into account privacy and confidentiality issues, which are often the main barriers to data sharing.
- *Dynamic Template* - In order to achieve data integration without privacy exposure, a data skeleton template has to be built for each data source. The template schema of the skeleton can be defined for each entity or can be reused for various entities, in the case of data aggregation. The most common types of entities used in biomedical sciences are Electronic Health Records (EHR), Electronic Data Capture (EDC), cohorts, medical imaging repositories and observational databases, among others. The architecture of the system should allow uses to easily create their data skeleton or a skeleton template, with the support of daily tools, such as a spreadsheet, for example.
- *Access Control* - Data privacy goes hand in hand with giving data owners fine-grained access control over their data. This can be achieved by creating different user profiles. The system should allow system administrators to dynamically modify or update any template schema. Data providers should be allowed to edit the information of their data's skeleton.
- *Data Discovery* - In their first interaction with the system, most users are interested in finding data entities aligned with their research interests. Therefore, a search functionality should be integrated within the system.
- *Data Comparison* - Comparing data belonging to different data entities should be enabled in the system.
- *Data Query* - After identifying the data entities, querying the data that they present within the system should be possible. While real data cannot flow out of healthcare institutions boundaries, users should be able to query the high-level information that has been integrated within the system.
- *Data Aggregation* - Labeling or grouping within the same structure data entities that concern the same topic should be possible.
- *Data Statistics* - Extracting statistical measurements from the integrated data is another important requirement. Not only the capability of having a global statistics dashboard should be featured, but also the ability to extract statistical measurements for individual data entities.

– *User-friendly Dashboard* - The system should have a web dashboard with global information about the data sources that are integrated.
– *Security Strategy* - Last but not least, a security strategy should be applied and Single Sign On (SSO) implemented over the various components of the system.

Several other non-functional requirements can also be defined, namely modularity and portability.

3.2 Motivational Example

The architecture of the system that we propose in this paper was designed for centralizing and sharing public or private data, whatever the data model or its purpose. Due to its architecture and flexibility, MONTRA can be applied to gather any type of data. However, the work that we present here was motivated by the plethora of geographically scattered biomedical datasets and the understanding of how life sciences could evolve if such datasets could be reused.

There has been an exponential increase in the volume of clinical and disease-specific data throughout Europe over the last few years, and there is now a need to link these data to provide additional benefits. Even though a considerable amount of relevant patient health information does exist, it is usually contained in a variety of systems stored in different locations, which inhibits efficient access from a central place. It is very important for researchers to have access to related data, and this work presents an effort towards accomplishing the integration of data across multiple heterogeneous data sources.

Generically, our research motivation was based on how to gather and integrate knowledge from M different types of registries among N different healthcare units. Our goal was to create a software solution that facilitates the integration and aggregation of biomedical data, enabling data discovery and promoting data sharing without breaking privacy rules. A basic usage scenario of MONTRA shows two different views of the motivation behind it. On one hand, we have a clinical researcher that needs to identify datasets that could support a study that (s)he intends to conduct. Instead of contacting different healthcare units and data custodians in order to find an answer, the researcher can access MONTRA and through its user-friendly interface find and compare a list of data entities that match given search criteria. The second view is that of a data custodian who wants to contribute to the advancement of biomedical sciences and is in need of a way to share the data that he or she is responsible for, without disclosing private or confidential information. MONTRA supports the data custodian by providing a flexible solution for building the skeleton of the data and sharing it without exposing patients' private data.

4 Results

4.1 System Architecture

After having identified the existent lacuna in the process of integrating biomedical data sources and having gathered requirements from end-users, we constructed a flexible architecture for centralizing and sharing biomedical data coming from multiple, independent, heterogeneous sources, as well as a user-friendly interface allowing interaction with the data. The data sources that are characterized within the system contain the full data, while the dashboard available through MONTRA provides an integrated view of the skeleton of these underlying sources.

The heterogeneous data sources can be of any type, from EHR and EDC records, cohorts or even patient files. The use of a dynamic data skeleton for data characterization and integration represents a layer of abstraction that does not depend on data models or physical data supports. The core of our approach, which we refer to as the skeleton, is represented in Fig. 1.

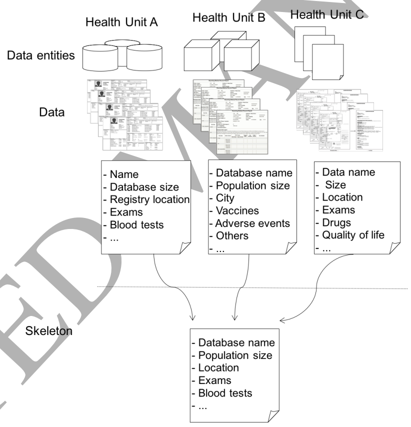


Fig. 1. The process of building a data skeleton. Data coming from different health units, or, more generally put, different data entities can be defined based on different structures. A data skeleton is a collection of metadata that can describe, in the same way, the information that exists in each of the initially different data entities. By complying to this skeleton, the data entities will expose the same information, which leads to data harmonization.

The skeleton represents a collection of metadata that best encapsulate the real data, which has to remain private. The skeleton definition is a straightforward operation that can be done by any data custodian and can be saved as a spreadsheet file and submitted through the web interface. If data aggregation is intended, the same skeleton template can be reused for the characterization of multiple data sources.

Different data entities complying to the same skeleton template will have a common representation within the platform. Each data source, that might belong to different institutions, will be characterized by the same fields, which, in a more general way, represent answers to the same questions defined in the skeleton template. By using a common structure as a metadata skeleton, distinct data sources converge to a homogeneous exposure of their metadata.

A simplified overview of MONTRA is summarized in Fig. 2. The metadata encapsulated in the skeleton of a data entity is also referred to in the system as a database. Users of MONTRA can browse, search, compare and query databases listed in the web interface.

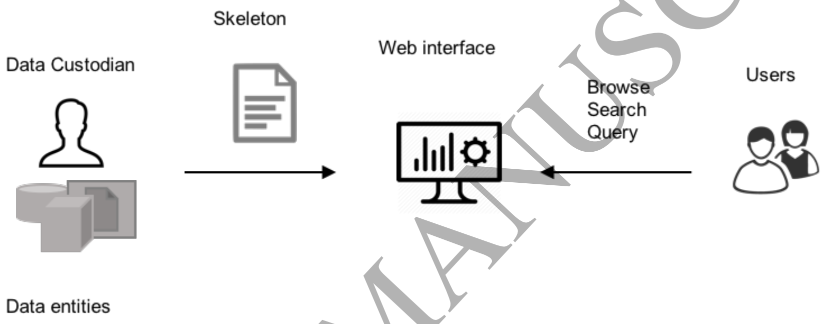


Fig. 2. MONTRA General View.

MONTRA was planned as a Rapid Application Development (RAD) [34] system, to allow fast deployment for a specific use case, and by following the Agile practices in software development [35]. RAD systems, unlike conventional ones, evolve as the project advances. They do not rely on rigid initial specifications, but are rather continuously adjustable in order to fit new requirements that arise as the project progresses and new knowledge is generated.

4.2 Data Skeleton

To integrate distinct types of data entities from different sources, we created a system that supports several schemas. Taking into account that the broader concept of skeleton is a set of aggregated data about a data entity, it could be translated into questionnaires or aggregated data that could be imported in CSV files. Thinking of the skeleton as a questionnaire, we have implemented a schema that contains questions and answers about one or more data entities. The key idea behind the skeleton schema is that we do not know what kind of questions or data will be introduced. Thus, our implementation needs to be flexible and not bound to a fixed set of questions. To accomplish this, we developed a dynamic questionnaire with multiple groups of questions, which can have dependencies between them. The questionnaire represents the type of data entity that will be skeletonized. Each questionnaire is composed of groups of questions. A group of questions is denominated a QuestionSet. Then, each QuestionSet is formed of Questions that can be of any type. The implementation of the answers to these questions is very flexible and extensible. Each skeleton schema contains several QuestionSets. Within these, each question can be individually defined, as well as the answer type - Table 1.

Table 1. Question types and the respective rendering.

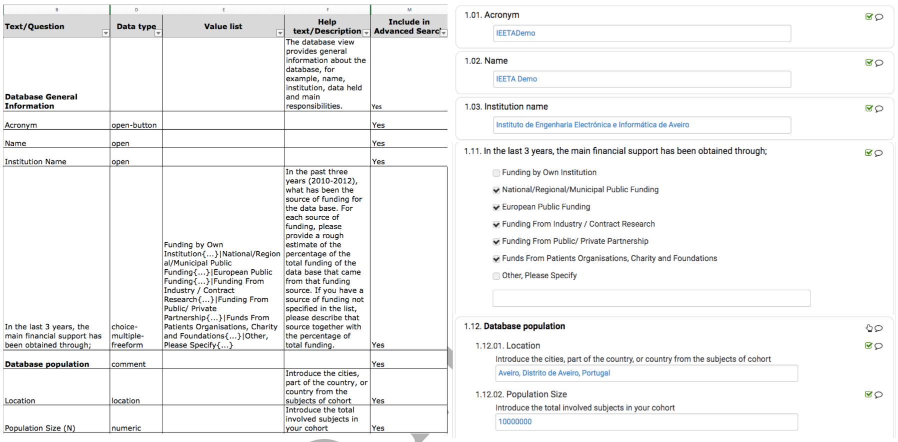
Component	**Visual render**
open	Open Answer, single line [input]
open-button	Open Answer, single line [input] with a button to validate
open-textfield	Open Answer, multi-line [textarea]
choice-yesno	Yes/No Choice [radio]
choice-yesnocomment	Yes/No Choice with optional comment [radio, input]
choice-yesnodontknow	Yes/No/Don't know Choice [radio]
comment	Comment Only
choice	Choice [radio]
choice-freeform	Choice with a freeform option [radio]
choice-multiple	Multiple-Choice, Multiple-Answers [checkbox]
choice-multiple-freeform	Multiple-Choice, Multiple-Answers, plus freeform [checkbox, input]
publication	Publication
datepicker	Date choice

The skeleton schema can be developed using a common spreadsheet to easily reach the end-users, usually data owners. If the skeleton is intended as a reusable template, the skeleton does not include the answers to the questions that have been defined in it. The skeleton is uploaded once to create the catalogue template, and the questions are rendered as shown in Figure 3. Each data custodian will fill in online the respective database or other entity information. If the skeleton is intended for just one database, the answers can be included and the complete spreadsheet can be uploaded through the web interface. In both situations, the information defined in the skeleton can be edited at any time. An import service is available to automatically load both a reusable skeleton template or a filled-in skeleton.

Some of them are generic, open text fields or numeric ones, while others allow the users to restrict the answer to a question to a specific type. Such fields are, for example, the date field, publication field or the location. A date field is visually rendered as a calendar drop-down that eases the user's task of typing a specific date. Similarly, locations can be chosen from a dynamic drop-down list that includes countries, regions, districts and cities.

The Publication question, as named in Table 1 is based on a web widget designed for this purpose. Data custodians can simply add one or a list of Pubmed ids and the widget fetches the information from the scientific publications identified by those ids. It does not only fetch the title and metadata about the publication, but also its abstract. Thereafter, the abstract is annotated using Becas [36], a web application service that provides biomedical concept identification. With this annotation, the user will have a better notion of the concepts identified in the database publication and will be linked to other relevant knowledge resources.

We intended MONTRA to be as open and standardless as possible. However, we maintain the possibility for the end-user to restrict the answer to a set of pre-defined values or answer types.



Text/Question	Data type	Value list	Help text/Description	Include in Advanced Searc
Database General Information			The database view provides general information about the database, for example, name, institution, data held and main responsibilities.	Yes
Acronym	open-button			Yes
Name	open			Yes
Institution Name	open			Yes
In the last 3 years, the main financial support has been obtained through;	choice-multiple-freeform	Funding by Own Institution{...}|National/Regional/Municipal Public Funding{...}|European Public Funding{...}|Funding From Industry / Contract Research{...}|Funding From Public/ Private Partnership{...}|Funds From Patients Organisations, Charity and Foundations{...}|Other, Please Specify{...}	In the past three years (2010-2012), what has been the source of funding for the data base. For each source of funding, please provide a rough estimate of the percentage of the total funding of the data base that came from that funding source. If you have a source of funding not specified in the list, please describe that source together with the percentage of total funding.	Yes
Database population	comment			Yes
Location	location		Introduce the cities, part of the country, or country from the subjects of cohort	Yes
Population Size (N)	numeric		Introduce the total involved subjects in your cohort	Yes

1.01. Acronym

IEETADemo

1.02. Name

IEETA Demo

1.03. Institution name

Instituto de Engenharia Electrónica e Informática de Aveiro

1.11. In the last 3 years, the main financial support has been obtained through;

□ Funding by Own Institution
☑ National/Regional/Municipal Public Funding
☑ European Public Funding
☑ Funding From Industry / Contract Research
☑ Funding From Public/ Private Partnership
☑ Funds From Patients Organisations, Charity and Foundations
□ Other, Please Specify

1.12. **Database population**

1.12.01. Location

Introduce the cities, part of the country, or country from the subjects of cohort

Aveiro, Distrito de Aveiro, Portugal

1.12.02. Population Size

Introduce the total involved subjects in your cohort

10000000

Fig. 3. Mapping from the skeleton definition into HTML5 forms. On the left, a view of some fields from the skeleton template (open-text, multiple choice, numeric and location). On the right, their rendering on the web interface.

4.3 CRUD Operations

Within software engineering, the acronym CRUD stands for Create, Read, Update and Delete entities. These operations are the four basic functions that are provided by relational database applications. Within the data catalogue created with a particular schema based on MONTRA, CRUD operations are automatically available for registered users, i.e. they can create, read, update and delete their own registers. A data custodian defines the data skeleton and fills in the respective information in an Excel file, as described above. After uploading the file through the web interface, its registry is displayed, similar to the information shown in Figure 3. Furthermore, the user interface allows viewing, searching, modifying and deleting information through computer-based forms.

With the registry online, the data custodian can decide if the registry is submitted as a draft or as public. Draft registries are only visible to the data custodian. At any time he or she can edit the information that was filled in. A registry can have more than one administrator. By default, the register owner is user submitting it, but ownership and administration privileges can be shared with other users. In addition to these operations, a private link of a registry can be created and shared with non-registered users. The web interface generates a private link which can be sent directly from it to an external user. The non-registered user will only have browsing privileges for that particular registry.

4.4 Searching Services

Within the system, users can have access to all the catalogue entries, according to their user profile. They are able to browse the registers but also to search for specific free text terms. A second search feature, which we called advanced search, is also available. This feature allows users to specify a more fine-grained search following the skeleton schema.

To support the searching backend, all the questions and answers are indexed by Solr[3]. A retrieval model was also built to score, sort and improve the quality of results. For instance, in a yes/no question such as "BMI measurements?", if the answer for a particular register is "yes", then, if a query by "BMI" is made, this entry should be retrieved. This means that not only the answer should be analyzed, but also the question, according to its type. The developed backend was also used to give suggestions in the free text search. The autocomplete suggestions are supported by another core (index schema) of Solr. For implementation, we rely on a tokenizer with an edge filter n-gram between 1 and 25. This is particularly useful to accelerate the process of giving accurate suggestions while the user is typing the text.

Since the MONTRA architecture is data-agnostic, i.e. it does not assume any particular end-user application, the Search functionality was designed as a flexible retrieving approach, to maximize the recall over the precision.

Additional complexity can be provided by the advanced search by using boolean logic terms. When this type of search is performed a boolean query is created, e.g. a query that includes a relation between two or more concepts. With these queries, the user can combine multiple search criteria to search for specific terms in each question that comprises the registry skeleton. The search engine understands the combination of AND and OR terms to filter content that will be displayed in the search results (Figure 4). Additionally, the user can save boolean queries within his account and retrieve and reuse them at any time.

Apart from retrieving and searching for terms of interest, it is also possible to compare several skeletons, taking into consideration a similarity metric based on the Levenshtein distance [37] for textual information, and on the cosine similarity [38] between records. The comparison feature enables the user to identify metadata similarities across distinct data entities. The results highlight in red

[3] http://lucene.apache.org/solr/

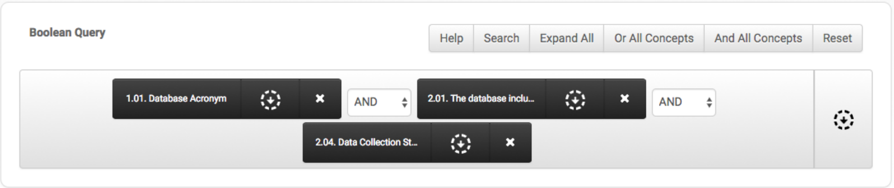
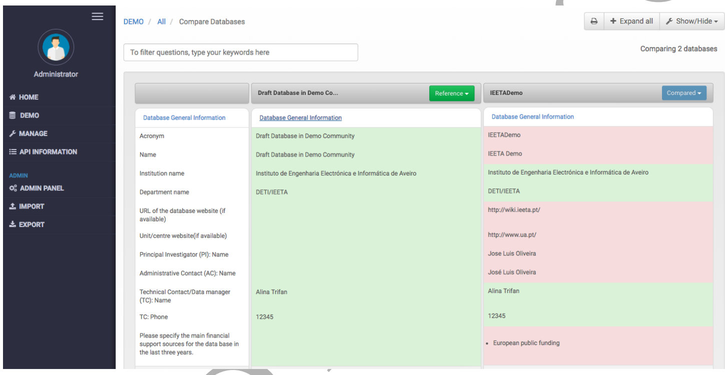


Fig. 4. An example of a boolean query that can be performed in the advanced search.

the information that varies, while the information that is common appears in green (Figure 5).



	Draft Database in Demo Co...	IEETADemo
Database General Information	Database General Information	Database General Information
Acronym	Draft Database in Demo Community	IEETADemo
Name	Draft Database in Demo Community	IEETA Demo
Institution name	Instituto de Engenharia Electrónica e Informática de Aveiro	Instituto de Engenharia Electrónica e Informática de Aveiro
Department name	DETI/IEETA	DETI/IEETA
URL of the database website (if available)		http://wiki.ieeta.pt/
Unit/centre website(if available)		http://www.ua.pt/
Principal Investigator (PI): Name		Jose Luis Oliveira
Administrative Contact (AC): Name		José Luis Oliveira
Technical Contact/Data manager (TC): Name	Alina Trifan	Alina Trifan
TC: Phone	12345	12345
Please specify the main financial support sources for the data base in the last three years.		• European public funding

Fig. 5. Database comparison. Similarities to the left-side data entity are shown in green, while the differences appear in red.

4.5 Plugins Integration

MONTRA supports third-party components and allows users, with adequate permissions, to extend its functionalities without having to deal with its base code. To accomplish this strategy, a microkernel architecture was developed, providing the platform's core to which several components can be dynamically added, enriching the overall system's functionality.

Two different types of components, or plugins, can be added to the platform by third-party developers:

- *Global*: they provide general services for a MONTRA instance. Once added to the platform, these plugins will be available on the user dashboard and/or in the main menu.

– *Registry*: this type of plugins provides added functionalities for each data record, and as such will be available as new services over the skeleton.

These plugins can be further divided into two types: third-party plugins and fully-fledged plugins. Third-party plugins are full web applications that are linked to the system, through the navigation menu. They usually provide a completely different functionality. The main goal of these plugins is to integrate their application features in MONTRA's environment. Fully-fledged plugins are internal extensions and they provide additional data services within the system. This type of plugin allows the development of decoupled services that can be easily attached to the main core module.

To allow its integration in the same web interface, each plugin is supported by a re-rendering method (Figure 6). After its initial representation, in the browser, each time an event occurs the plugin updates its content using *self.html()* or *self.append()*. When this is done, *self.refresh()* has to be executed to update the visualization.

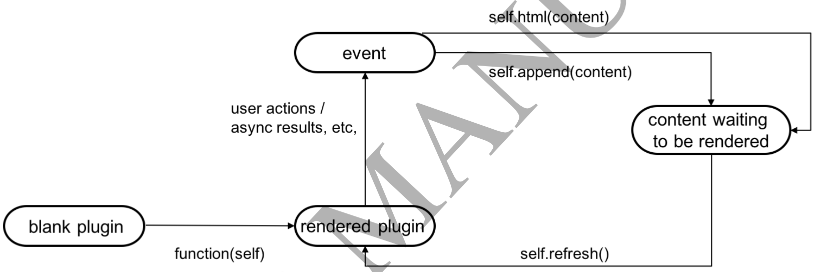


Fig. 6. Plugin rendering lifecycle.

The plugin development process follows a specific lifecycle that is managed internally within MONTRA services (Figure 7). The first step is to obtain the administrator's approval to develop plugins. After that, the user can create any number of plugins. Each plugin can have a series of versions, although users can develop and live-preview their plugin versions during development. Whenever a plugin version is deemed ready for production, it must be submitted for the administrator's approval to become available. Any further changes to an already approved version will remove the approval status, and the plugin will have to be submitted again. This workflow assures the quality control of components added by external developers.

Through this plugin functionality, a MONTRA-based platform can be extended with external components, which may provide, beyond the original metadata, complementary information of the data entities, such as aggregated data or summarized views. Such dashboards can include, for instance, the number of patients per age and per year, the average time of follow-up and many other

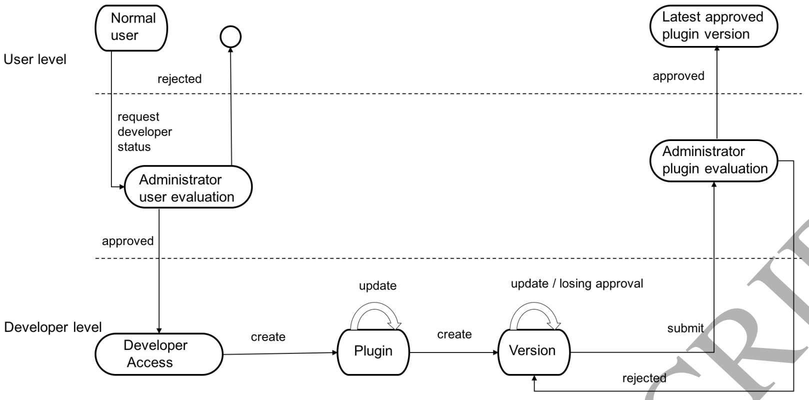


Fig. 7. Plugin development lifecycle.

statistics. Taking the example of EHR data, a population characteristics dashboard can be integrated as a third-party plugin (Figure 8).

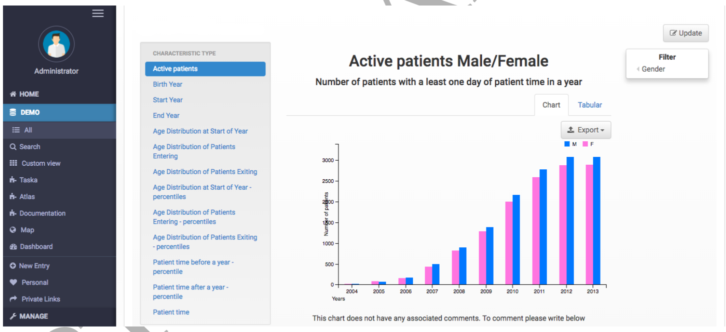

Administrator
HOME
DEMO
All
Search
Custom view
Taska
Atlas
Documentation
Map
Dashboard
New Entry
Personal
Private Links
MANAGE
CHARACTERISTIC TYPE
Active patients
Birth Year
Start Year
End Year
Age Distribution at Start of Year
Age Distribution of Patients Entering
Age Distribution of Patients Exiting
Age Distribution at Start of Year - percentiles
Age Distribution of Patients Entering - percentiles
Age Distribution of Patients Exiting - percentiles
Patient time before a year - percentile
Patient time after a year - percentile
Patient time
Update
Filter
Gender
Active patients Male/Female
Number of patients with a least one day of patient time in a year
Chart
Tabular
Export
M
F
Number of patients
0
500
1000
1500
2000
2500
3000
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
Years
This chart does not have any associated comments. To comment please write below

Fig. 8. Example of a third-party plugin integrated within a possible MONTRA-based platform. The plugin is a population characteristics dashboard that shows statistic information about the data entity, in this example, EHRs.

4.6 Role-Based Access Control

A Role-Based Access Control (RBAC) system is included in MONTRA to guarantee that proper access constraints are in place. Besides the users, groups and roles can be created to define permissions to access data and services.

A regular User can register in the system and, once accepted, s/he can browse the list of existing records. This user profile has access to the information available in the catalogue, e.g. can search, query, compare or export data.

A data Owner is a regular user who, at some point, added new entries into the catalogue. By entering data, that user is then responsible for the management and access control of those entries (can edit, decide about public/private permissions, share owner responsibilities, private share, ...).

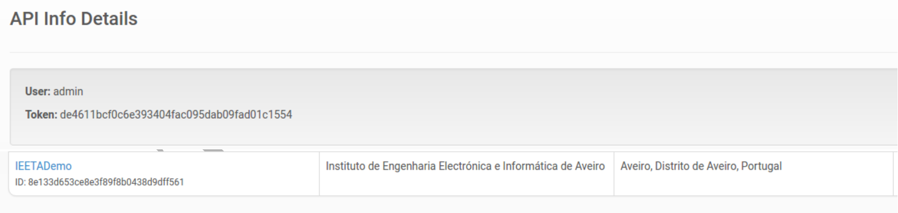
The system Administrator is able to create and edit skeleton schemas, validate user registrations, and the overall management of the platform.

A Developer is a user who is allowed to create several types of new components, by following the plugins workflow.

4.7 RESTful API

MONTRA integrates a RESTful API which provides a set of programmatic endpoints that can be consulted by third party applications. The main idea behind the Web API is that other applications can send data to MONTRA, in the format of key-value pairs, containing extra metadata of the registry. It is a simple mechanism to dynamically add metadata information in each entry.

A double key schema allows controlling the access to this API. To use a web service, third-party applications need to know two distinct tokens: the user token and the registry key. The access to the information is granted through the combination of both keys, being then possible to send extra data to the specific registry IDs. These tokens are available within MONTRA's web interface, in the user workspace. (Fig. 9).



API Info Details

User: admin

Token: de4611bcf0c6e393404fac095dab09fad01c1554

IEETADemo ID: 8e133d653ce8e3f89f8b0438d9dff561	Instituto de Engenharia Electrónica e Informática de Aveiro	Aveiro, Distrito de Aveiro, Portugal

Fig. 9. User token and registry id key.

The web service accepts information using key-value pairs, and associates it to the respective registry. An example of a JSON format that the API can receive:

```
{
 "registryID":"<your ID>",
 "values":
    {
            ”field1”:”value1”,
            ”field2”:”value2”,
            ”field3”:”value4”,
            ...
        }
}
```

4.8 Software Technologies

MONTRA is written in Python 2.7.6[4], using Django 1.4.5[5], a framework that encourages rapid development and clean programmatic design. However, a considerable part of the development was made in HTML5, CSS and JavaScript, namely the interface and the end-user interaction. Furthermore, in order to improve the web design quality, we have adopted the Bootstrap2[6] framework, a front-end framework for web development. To assure the system's performance when dealing with concurrent requests a local cache is maintained, based on memcached3[7]. Moreover, several tasks, such as indexing the skeleton in Solr, can take too much time, making users wait for the operation to complete. To manage these tasks, we rely on a queue message system (Rabbit MQ4[8]) and on the Celery[9] backend to execute them in the background.

5 Discussion

The solution that we propose for integrating a vast range of biomedical data coming from different, disperse institutions resides on using the same skeleton template for the same family of data sources (e.g. EHR repositories, disease-specific cohorts, etc.). A group, or a collection of data sources or, more generically, data entities, that are characterized by the same skeleton template will naturally converge toward a common structure in what concerns the information that it exposes.

Multiple similar data sources can be aggregated, or simply put, grouped as an entity that respects the same metadata skeleton and that is managed by one or multiple ”skeleton managers“. The manager can accept or reject the publishing of a data source or even request for changes in answers submitted by a given data custodian. The task of designing the skeleton template most certainly involves the collaboration of data custodians, but once having the skeleton decided, the process of filling in the data is straightforward. The effort spent into this task is not too significant compared to the final result of having a data source discoverable, shareable and overall useful for the research community.

[4] www.python.org

[5] http://www.djangoproject.com

[6] http://bootstrap.com

[7] http://www.memcached.org

[8] http://www.rabbitmq.org

[9] http://www.celeryproject.org/

MONTRA has the potential to simplify the setup of web-based catalogues, for many distinct scenarios and applications.

We have been applying this framework to build the EMIF Catalogue[10], [27] an online platform that aims to be a marketplace where data custodians can publish and share different levels of information about their clinical databases, while biomedical researchers can search for databases that fulfill their particular study requirements. In this scenario, the "Database" is the main entity characterized, typically EHR or cohorts. Currently, the EMIF Catalogue supports several distinct projects, combining, for instance, data available in pan-European EHR and Alzheimer cohorts. Currently, the EMIF Platform integrates 374 distinct databases and around 700 users.

Another example where MONTRA is being used is in maintaining a catalogue of neuroradiology clinical cases[11]. In this portal, which describes patient information, several plugins were added to include also the search and visualization of a Picture Archiving and Communication System (PACS) archive. This repository includes interesting case studies from the Portuguese Society of Neuroradiology. It is an annotated Web-based medical imaging repository used for clinical research and academic purposes.

A third scenario, currently being deployed, is to support the gathering and maintenance of Case Report Forms (CRF) together with omics data, for a cohort of patients suffering from heart failure with preserved injection.

6 Conclusions

The main motivation for this work was to facilitate the setup of web data catalogues for distinct applications. MONTRA, the presented system, is based on dynamic skeletons which allow describing any kind of data, being automatically used to create the data store and to build the web user interface, without the need to create a single line of code. This framework is being used and validated in several applications, such as the EMIF European project, to allow the presentation, discovery and share of biomedical data sources.

Acknowledgements

This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF grant n. 115372), and from the NETDIAMOND project (POCI-01-0145-FEDER-016385). We would like to acknowledge all of the team members that have contributed to the development of this architecture along the time. A list of the main contributors can be found here: `https://github.com/bioinformatics-ua/montra/graphs/contributors`.

[10] https://emif-catalogue.eu

[11] http://bioinformatics.ua.pt/spnr/

References

1. Alon Halevy, Anand Rajaraman, and Joann Ordille. Data Integration: The Teenage Years. In *Proceedings of the 32nd International Conference on Very Large Data Bases*, pages 9–16. VLDB Endowment, 2006.
2. Vasileios Lapatas, Michalis Stefanidakis, Rafael C. Jimenez, Allegra Via, and Maria Victoria Schneider. Data Integration in Biological Research: an Overview. *Journal of Biological Research-Thessaloniki*, 22(1):9, Sep 2015.
3. Pedro Lopes, Luis Bastião Silva, and José Luis Oliveira. Challenges and Opportunities for Exploring Patient-Level Data. *BioMed Research International*, 2015.
4. Tempest A. van Schaik, Nadezda V. Kovalevskaya, Elena Protopapas, Hamza Wahid, and Fiona G.G. Nielsen. The Need to Redefine Genomic Data Sharing: A Focus on Data Accessibility. *Applied & Translational Genomics*, 3(4):100 – 104, 2014. Global Sharing of Genomic Knowledge in a Free Market.
5. Douglas Teodoro, Rmy Choquet, Emilie Pasche, Julien Gobeill, Christel Daniel, Patrick Ruch, and Christian Lovis. Biomedical Data Management: a Proposal Framework. *Studies in Health Technology and Informatics*, 150:175179, 2009.
6. Maurizio Lenzerini. Data Integration: A Theoretical Perspective. In *Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 233–246, New York, NY, USA, 2002. ACM.
7. Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. *Nature Biotechnology*, 25(11):1251–5, 2007.
8. J Marc Overhage, Patrick B Ryan, Christian G Reich, Abraham G Hartzema, and Paul E Stang. Validation of a Common Data Model for Active Safety Surveillance Research. *Journal of the American Medical Informatics Association*, 19(1):54–60, 2011.
9. George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Studies in Health Technology and Informatics*, 216:574, 2015.
10. Ivan Merelli, Horacio Pérez-Sánchez, Sandra Gesing, and Daniele DAgostino. Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives. *BioMed Research International*, 2014.
11. P. Beynon-Davies, C. Carne, H. Mackay, and D. Tudhope. Rapid Application Development (RAD): An Empirical Review. *European Journal of Information Systems*, 8(3):211–223, 1999.
12. Marco Masseroli, Barend Mons, Erik Bongcam-Rudloff, Stefano Ceri, Alexander Kel, François Rechenmann, Frederique Lisacek, and Paolo Romano. Integrated Bio-Search: Challenges and Trends for the Integration, Search and Comprehensive Processing of Biological Information. *BMC Bioinformatics*, 15(1):S2, Jan 2014.
13. Walter Sujansky. Heterogeneous Database Integration in Biomedicine. *Journal of Biomedical Informatics*, 34(4):285 – 298, 2001.
14. The Global Alzheimer's Association Interactive Network (GAAIN). *Alzheimer's & Dementia*, 11(7):P121, 2015.

15. Shawn N Murphy, Michael E Mendis, David A Berkowitz, Isaac Kohane, and Henry C Chueh. Integration of Clinical and Genetic Data in the i2b2 Architecture. In *AMIA Annual Symposium Proceedings*, volume 2006, page 1040. American Medical Informatics Association, 2006.
16. William Hersh and Stephen Wu. IR Meets EHR: A Patient Cohort Discovery Task.
17. Genomics Cohort Catalogue. https://genomics.uq.edu.au/cohorts. Accessed: 2017-07-26.
18. Isabelle Budin-Ljøsne, Julia Isaeva, Bartha Maria Knoppers, Anne Marie Tassé, Huei-yi Shen, Mark I McCarthy, Jennifer R Harris, ENGAGE Consortium, et al. Data Sharing in Large Research Consortia: Experiences and Recommendations from ENGAGE. *European Journal of Human Genetics*, 22(3):317, 2014.
19. Owen Lancaster, Tim Beck, David Atlan, Morris Swertz, Dhiwagaran Thangavelu, Colin Veal, Raymond Dalgleish, and Anthony J Brookes. Cafe Variome: General-Purpose Software for Making GenotypePhenotype Data Discoverable in Restricted or Open Access Contexts. *Human Mutation*, 36(10):957–964, 2015.
20. CLOSER DISCOVERY-A Resource for Social Science Researchers using Longitudinal Data. http://www.closer.ac.uk/event/intro-closer-discovery-bristol/. Accessed: 2017-07-26.
21. Isabel Fortier, Parminder Raina, Edwin R Van den Heuvel, Lauren E Griffith, Camille Craig, Matilda Saliba, Dany Doiron, Ronald P Stolk, Bartha M Knoppers, Vincent Ferretti, et al. Maelstrom Research Guidelines for Rigorous Retrospective Data Harmonization. *International Journal of Epidemiology*, 46(1):103–105, 2017.
22. Deborah A Zarin, Tony Tse, Rebecca J Williams, Robert M Califf, and Nicholas C Ide. The clinicaltrials. gov results databaseupdate and key issues. *New England Journal of Medicine*, 364(9):852–860, 2011.
23. Sanchita Bhattacharya, Sandra Andorf, Linda Gomes, Patrick Dunn, Henry Schaefer, Joan Pontius, Patty Berger, Vince Desborough, Tom Smith, John Campbell, et al. Immport: disseminating data to the public for the future of immunology. *Immunologic research*, 58(2-3):234–239, 2014.
24. Chao Pang, David van Enckevort, Mark de Haan, Fleur Kelpin, Jonathan Jetten, Dennis Hendriksen, Tommy de Boer, Bart Charbon, Erwin Winder, K Joeri van der Velde, et al. Molgenis/connect: a System for Semi-Automatic Integration of Heterogeneous Phenotype Data with Applications in Biobanks. *Bioinformatics*, 32(14):2176–2183, 2016.
25. Christopher J Mungall, Julie A McMurry, Sebastian Köhler, James P Balhoff, Charles Borromeo, Matthew Brush, Seth Carbon, Tom Conlin, Nathan Dunn, Mark Engelstad, et al. The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic acids research*, 45(D1):D712–D722, 2016.
26. Sharon F Terry. The Global Alliance for Genomics & Health. *Genetic Testing and Molecular Biomarkers*, 18(6):375–376, 2014.
27. EMIF - European Medical Information Platform. http://www.emif.eu/. Accessed: 2017-07-26.
28. Lindsey C Crosswell and Janet M Thornton. Elixir: a distributed infrastructure for european biological data. *Trends in biotechnology*, 30(5):241–242, 2012.
29. Ana Marta Anes, Alejandro Arana, Kevin Blake, Jan Bonhoeffer, Stephen Evans, Annie Fourrier-Réglat, Jesper Hallas, Ursula Kirchmayer, Victor Kiri, Olaf Klungel, et al. The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). Guide on Methodological Standards in Pharmacoepidemiology (2014). 2012.

30. Bridge to data. https://www.bridgetodata.org/. Accessed: 2017-06-16.
31. HCUP Databases. Agency for Healthcare Research and Quality. *Rockville, MD*, 2014.
32. Procurement of Shared Data Instruments for Research Electronic Data Capture (REDCap). *Journal of Biomedical Informatics*, 46(2):259 – 265, 2013.
33. Paul A. Harris, Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, and Jose G. Conde. Research Electronic Data Capture (REDCap)-a Metadata-driven Methodology and Workflow Process for Providing Translational Research Informatics Support. *Journal of Biomedical Informatics*, 42(2):377–381, April 2009.
34. James Martin. *Rapid Application Development.* Macmillan Publishing Co., Inc., Indianapolis, IN, USA, 1991.
35. Luis Bastiao Silva, Rafael C Jimenez, Niklas Blomberg, and José Luis Oliveira. General guidelines for biomedical software development. *F1000Research*, 6, 2017.
36. Tiago Nunes, David Campos, Sérgio Matos, and José Luís Oliveira. Becas: Biomedical Concept Recognition Services and Visualization. *Bioinformatics*, 29(15):1915–1916, 2013.
37. Michael Gilleland et al. Levenshtein distance, in three flavors. *Merriam Park Software: http://www. merriampark. com/ld. htm*, 2009.
38. Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.