MAGIC: Manuscripts of Girolamini in Cloud

IOP Conf. Series: Materials Science and Engineering **949** (2020) 012081 doi:10.1088/1757-899X/949/1/012081

MAGIC: Manuscripts of Girolamini in Cloud

Guido Russo[(1),(2),(11)], Luciano Aiosa[(3)], Giancarlo Alfano[(1)], Angelo Chianese[(1)], Fabio Cornevilli[(4)], Gian Marco Di Domenico[(5)] Pasqualino Maddalena[(1)], Andrea Mazzucchi[(1)], Ciro Muraglia[(6)], Felice Russillo[(7)], Alessandro Salvi[(8)], Bernardino Spisso[(9)], Guido Trombetti[(10)], Giuseppe Zollo[(1)]

(1) Università degli studi di Napoli Federico II, Italy – email *name.familyname@unina.it*
(2) Istituto Nazionale di Fisica Nucleare, Napoli, Italy – email *guido.russo@unina.it*
(3) SA Lombardia, Milano, Italy – email *l.aiosa@consorziocsa.it*
(4) System Management, Italy – email *fcornevilli@sysmanagement.it*
(5) SA Documents, Roma, Italy – email *gm.didomenico@consorziocsa.it*
(6) Netcom Group S.P.A., Napoli, Italy – email *c.muraglia@netcomgroup.eu*
(7) Databenc s.c.a.r.l., Napoli, Italy – email *felice.russillo@gmail.com*
(8) Netcom Group S.p.A., Napoli, Italy – email *a.salvi@netcomgroup.eu*
(9) Istituto Nazionale di Fisica Nucleare, Napoli, Italy – email *spisso@na.infn.it*
(10) Università degli studi di Napoli Federico II, Italy – email *guitrombetti@gmail.com*
(11) Author to whom any correspondence should be addressed: *guido.russo@unina.it*

Abstract. We describe a project, fully formalized, to create a "Service Center" for ancient books and manuscripts, in the "Biblioteca dei Girolamini" (Library of Girolamini) in Naples, Italy. This unique library hosts about 180.000 pieces, 500 of which are medieval manuscripts, 120 incunabulum, 5.000 editions of the 16th century – among others, as well as an ancient Musical Archive. The library is located in the heart of Naples, in a magnificent place where a series of initiative are already taking place, as a collaboration between the University of Naples Federico II and the Ministry of Cultural Heritage. All these features make Library of Girolamini a rare case study in the field of ancient books. The project aims at a complete digitization of the books; however, the important part of the project is the creation of full sets of metadata with a complete history of the documents and of the studies about the document itself. The first books to be considered will be a set of 512 manuscripts of the XIII century, e.g. a fully illustrated Divina Commedia or Seneca's tragedies with miniature paintings. There will be two main streams of work: a) conservation and study of ancient books through the digitization in a proper file format namely the FITS (**F**lexible **I**mage **T**ransport **S**ystem); b) grant the access to this information in the Big Data era using such technologies as Internet of Things and Machine Lerning in order to identify the ancient book and categorize it. This project will also allow to many young students to put in practice the studies within the "*Scuola di Alta Formazione in Storia e Filologia del manoscritto e del libro antico*", formally started in the year 2018. The access to information will allow, for example, the extraction and classification of figures, and the automatic recognition of the device used (pc, tablet, smartphone) and its resolution.

1. Introduction

Several library collections of ancient books existing in the world, and particularly in Europe, can have a new life and new possibilities of fruition, through a massive process of digitization. An example is the project *Digita Vaticana* [1] [2], which aims at digitalizing more than 80,000 manuscripts of the

Apostolic Library: started in 2012, the project has already reached about 23% of its goal, and the data are accessible through web in several paradigm.

Another example is the *MünchenerDigitalisierungsZentrum* (MDZ) in Germany, which handles the digitization and online publication of the cultural heritage preserved by the Bavarian State Library and by other institutions [3]. Nowadays, it provides one of the largest digital collections in Germany, including more than 900,000 titles available online and the work is still going on [4] [5].

The whole process of digitization of libraries around the world, after some pioneering work in the late 80's, started early in the 90's without a serious coordination. Although it was clear to all projects that the digitization itself was only a first step, and that the important part was how to make these digital data available to people, nevertheless several errors were made. Every project chose its own format for digital images, often adopting an existing open format, but in some cases defining a new proprietary format which prevented the interoperability of libraries and the exchange of data [6].

The problem in the data format is that one has to use a format which has to last several decades, accommodating for future, unknown enhancements, and that allows for different data resolution. Over decades, it may happen that some part of the file becomes unreadable, but the data format has to allow also in this case a partial file reading where most of the commonly used format do not allow this.

Another problem is the long-term archiving: while we can still read documents written hundreds of years ago, it is not clear how to maintain digital data for centuries.

In this paper we will address some of these technical aspects, and the core of the project we have setup as a collaboration between academic institutions and private companies.

2. The Library of Girolamini

2.1. The Library and its history

The Library of Girolamini is a rare episode in the history of books and culture in Italy and in the world. Carried out over the centuries through the subsequent composition of private funds, the Library is today preserved in the Girolamini complex, in this precious corner of the historic center of Naples, between Greek-Roman ruins, medieval churches and sumptuous seventeenth-century buildings along the Via dei Musei, a few steps from the Duomo, 200 meters from the new, futuristic metro station designed by Fuksas.

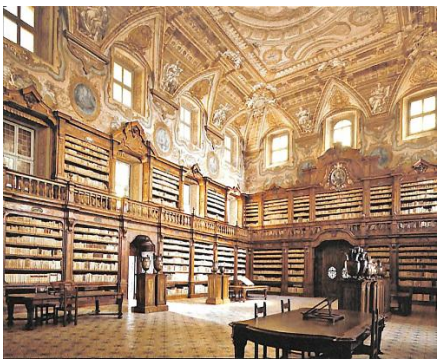

Figure 1. Library of Girolamini

The library is annexed to the National Monument of the Girolamini of Naples and it is focused in Christian theology, philosophy, Christian church in Europe, history of the church, sacred music and general history of Europe. Housed in a Girolamini Oratory, in opposition to the uses of the monastic orders, which did not admit the public into their libraries, the Institute was opened to the public in 1586. The Library is one of the richest in the South Italy. In 1727 the Oratorian fathers, on the advice of Vico, purchased the Library of Giuseppe Valletta which included a rich collection of legal, philosophical, religious and literary texts of the seventeenth and of the Neapolitan eighteenth century.

The Library depends from the Ministry of Cultural Heritage, more detailed info can be obtained at the site http://www.bibliotecadeigirolamini.beniculturali.it/.

2.2. The collection

The Library has a collection of books of about 159,700 units including volumes, pamphlets, 137 musical prints, 5,000 editions of the sixteenth century, 120 incunabula, 10,000 rare and valuable editions, 485 periodicals, a large quantity of microfilms and portraits. Several funds have enriched the institute's patrimony among which 5,057 volumes of the *Agostino Gervasio* Fund, whose texts deal with archeology, numismatics and classical literature, the *Filippino* Fund, mainly of ecclesiastical history, sacred writings and theology, the *Giuseppe Valletta* Fund, containing rare editions of the sixteenth and

seventeenth centuries consisting of Latin and Greek classics, history and philosophy, and the 940 volumes of the *Valeri* Fund concerning the history of Naples and southern Italy.

3. The MAGIC project

In 2019, a collaboration between private companies and a research institutions, has submitted a project in the filed of digital libraries. The project, named MAGIC (**MA**nuscripts of **GI**rolamini in **C**loud, is detailed in the following pages.

3.1. The goal

The project proposal has as the ultimate goal the creation of a Service Center for ancient books and manuscripts, considering as a first step an entire section in the Library of Girolamini. The Center is designed so that it can generate culture and create value, through concrete activities in the fields of restoration, digitization, cataloging, also forming young people in all these fields, who can exploit diversified methodologies and skills. A first goal will be the long-time preservation of the contents of the ancient books through the digitization using proper scanning and storage equipment (described below) and choosing an open file format namely the FITS (**F**lexible **I**mage **T**ransport **S**ystem) which grants (among other features) a great durability of the image *(*see sect. 5.1). Another main topic of the project is the smart access to this huge amount of information which will rely on a strong categorization of the contents and on technologies borrowed from the Big Data realm, such as Machine Learning (ML) and unstructured databases. Beside, the Internet of Things paradigm will be applied in order to physically locate the ancient document connected to a particular file.

The categorization process, of the handwritten testimonies, will highlight both the philological and critical aspects of the text and those more purely material. Ancient books and manuscripts will be studied and processed with reference to the specificity of the book object as an expression of historical, social and cultural instances connected to the era and to the environment.

The Service Center will focus on the different aspects that characterize the book: the history and tradition of the texts; the graphic features; material aspects (desk support, manufacturing techniques, ornamentation).

3.2. The process of digitization and archiving

The project team has identified a number of tasks and subtasks, which are schematically represented in figure 2, and which will be followed for any single manuscript of the Girolamini's collection.

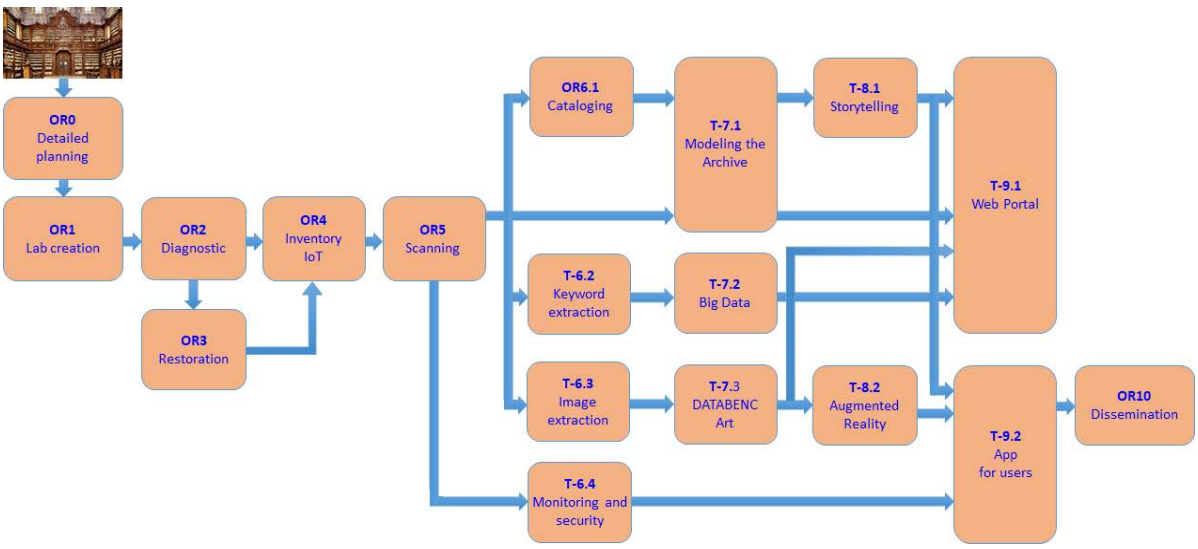

Figure 2. The process workflow for the activities within the MAGIC project

3.3. Details of the process workflow

OR0 – Detailed planning: The nature itself of the ancient documents needs an accurate planning of the whole process, in particular for what concerns the handling of the documents; this will be done also to obtain all the authorizations to move the documents, temporarily, from their original places.
OR1 – Lab creation: The construction of the lab, including the clean room later depicted in fig. 3, is a task which will be started among the first.
OR2 – Diagnostic: A diagnosis of each document (books, manuscript) is necessary in order to validate the feasibility of the scanning process, which could be in bad conditions that inhibit the manual handling. A first cataloguing will be done in this subtask, with respect to its physical parameters.
OR3 – Restoration: If the damage is not too critical, specialized technicians will start a restoration process without stopping the overall chain of digitization.
OR4 – Inventory-IoT: A second step in cataloguing will include a refined metadata creation and the creation of RFID (**R**adio-**F**requency **ID**entification) tags containing these information, under the logic described later in sec. 6.2.
OR5 – Scanning: The most time-consuming task, which will require special scanners (fig. 4 and 5 in the following) and accurate handling of the documents, in order to produce the FITS files of sect. 5.1.
OR6 – Cataloging, keyword and image extraction, monitoring: In parallel with scanning, people of the University will define a complete keyword set for each document to enhance the metadata database.
OR7 – Modeling and Big Data: The huge dataset which will come out from the digitization process will require an accurate modeling but also a Big Data approach; the University has a wide experience in this field, as it is already handling Petabytes of data coming from physics experiments and biotechnology genome databases. The Data Center of sect. 5.2 will host the archive, under the Cloud paradigm.
OR8 – Storytelling, augmented reality: before opening the archive to the public, a storytelling process will define the information to put online for each document to describe it, and will also experiment an augmented reality approach to selected items.
OR9 – Web Portal, App: Under this task we will create a modern web portal and smartphone Apps that will allow full access to the new archive of the Library of Girolamini.
OR10 – Dissemination: The whole project, and the availability of data, will be supported by a large information campaign, designed for both specialists and wide public.

4. The partners

4.1. SA Document

SA Documents srl is among the leading companies in Italy in the sector of Document Management solutions. Founded in 2010 by the desire to combine the great Italian archival tradition with the opportunities provided by modern ICT (**I**nformation and **C**ommunications **T**echnology), it aims to offer customers an integrated, simple and fast archiving and document management outsourcing service scientifically correct and attentive to the expectations.

4.2. SA Lombardia

SA Lombardia has a very strong specialization in data processing and management, as well as in the automation of authorization processes, through innovative solutions and systems produced by its own professionals and with its own equipment. Besides, SA Lombardia is a consortium partner of the company CSA S.c.a.r.l., which is one of the most important companies at national level in the field.

4.3. Università degli studi di Napoli Federico II

The University of Naples was named after Federico II to underline its very ancient origins, dating back to June 5 1224. Nowadays, the University has 20 Departments, about 90,000 students, about 2,000 professors and 4,000 technicians. Two of these departments have an active role in MAGIC: *i)* The Department of Humanities, which represents an aggregation of homogeneous knowledge within the humanities and social disciplines; *ii)* The Department of Physics, which has as its expertise in physical

analysis and ICT; it operates in strict collaboration with the Naples Section of INFN (**I**stituto **N**azionale di **F**isica **N**ucleare*)*.

4.4. NetCom Group

NetCom Group SpA operates in the ICT sector providing specialized consulting services in the field of software design, development, testing and validation for telco, automotive, aerospace, defense, health and public administration applications. The company can count on a high number of ICT specialists, to support companies in their daily activities and help the development of new products and services.

4.5. SCABEC

Scabec SpA was established by the Campania Region following the approval of Title V of the Constitution, which delegated the functions of valorisation and promotion of cultural assets to the Regions. The company, established to meet the general interest needs of the community, operates on the basis of its institutional mission. Therefore its reference market is represented by the local authorities of the territory as territorial emanations of the MIBAC (**MI**nistero per i **B**eni e le **A**ttività **C**ulturali).

*4.6. DATABENC (**D**istretto ad **A**lta **T**ecnologi**A** per i **BEN**i **C**ulturali)*

Although not a partner of the project, an important role in MAGIC is represented by the consulting services entrusted to DATABENC, which it is a consortium company established on October 2012. It was born from an idea promoted by the University "Federico II" and from the University of Salerno. DATABENC is an eco-system of open innovation, founded on an evolved network of public and private subjects (institutions, universities, research centers, etc.) that spreads and retains shareable value.

5. Technical aspects

In this section we describe the format chosen for data storing and the infrastructure planned (and already in costruction) for the physical storage.

5.1. The choice of the file format: FITS

Flexible Image Transport System (FITS) was initially developed by astronomers in the USA and Europe in the late 1970s to serve the interchange of data between observatories and was brought under the auspices of the International Astronomical Union in 1982 [7]. In 2019, FITS is still widespread as a data interchange and archiving format by a lot of scientists around the world due its architecture and Operating System (OS) independent file format designed to store, transmit, and manipulate scientific images and associated data. However, one has to consider the file content as data for analysis rather than simply as pictures to look at, and therefore FITS has a strong the metadata representation. A two-dimensional image with all the associated metadata, can be stored in a FITS file in a way which ensures long term preservation.

When FITS format is used for straightforward image data, i.e. the pages of a manuscript, the representation is typically a two-dimensional (X, Y) matrix with single values at each point (colour) or the colour can be encoded in a three-dimensional (X, Y, colour) matrix.

FITS was designed keeping in mind the long-term archival use and the maxim "once FITS, forever FITS" has guided the scientists to ensure that the format is backwards compatible as new features are added. Version 3 (the current one, published in 2012) includes support for 64-bit integers and for tables with variable-length arrays.

In addition to storing images that are extremely faithful to the original, a FITS file can contain many other information about the manuscript (dimensions, material ...); it is also free from legal restrictions and it is kept up to date through the largest international association of scientists in the sector, it is immune to viruses and can be viewed by any image processing program.

The whole Vatican Library is being archived in this FITS format, as it is believed to be the only existing image format capable of guarantee a long term preservation. When a page of the manuscript is

digitized, the image is acquired in JPEG2000 format and stored on a staging area. While a technicians continues to digitize other pages, another person adds metadata and creates the FITS file, and it is this file which is stored in the archive. All the metadata are within the file, and other information can also be added later on the same file, an essential feature of FITS, since the most important metadata will come from the researchers.

5.2. The physical infrastructure

The physical infrastructure for the MAGIC project is based on two different sites. The *first site* wil be within the Library of Girolamini, the reason for this choice lies in the need to minimize the ancient books physical stress and for safety reasons. The manuscripts have to be scanned in a room with controlled

Figure 3: The humidity-controlled room to be installed at Library of Girolamini.

Figure 4: A planetary scanner for use with ancient manuscripts

Figure 5: Scanning fragile bound books

humidity and temperature, to avoid damages to the paper which has now up to eight hundreds years. The technical choice is to install a clean room with humidity and temperature controlled, on the ground floor of the library room. The books will be entered into the room a few at a time until the complete digitization, which will use the so-called *planetary scanners* able to scan also the fragile bound documets like the ancient books.

The second site is located in the university complex of Monte Sant'Angelo, owned by University of Naples Federico II (http://www.unina.it). In this large academic site the University has built, with financial support form the MIUR (**M**inistero dell'**I**struzione, dell'**U**niversità e della **R**icerca), a Data Centre for scientific applications. The Data Centre has 33 racks, water refrigerated, each of them capable of up to 20 servers fully equipped. A part of these racks will accommodate the MAGIC project hardware. In particular, MAGIC will initially use 10 servers and 1 storage system with 1 Pbyte capacity, and this hardware is being installed at the moment of writing. The whole Data Centre is connected to the Internet through a 2 Gbit/sec optical link, fully symmetric, and the bandwidth will soon be upgraded to 10 Gbit/sec; this fast connect will allow access to the private cloud of the Data Centre to any user around the world. In 2019, a large financial support has been given to the Data Centre with a grant from PON 2014-2020 "Research Infrastructure", named project IBISCO – Infrastructure for Big data and Scientific Computing (project code PIR01_00011).

The Data Centre operates H24, 365 days a year, it is unattended but is fully controlled by a Control Room nearby, where all the almost 8,000 parameters that are continuously acquired in real time converge, to present a dashboard where an operator supersedes at the Data Centre efficiency.

6. Technology developments

MAGIC project proposal is characterized, from the point of view of technological developments, on three pillars that are increasingly characterizing the world of work: Big Data, Internet of Things, and Artificial Intelligence

Figure 6: the SCoPE Data Centre at the University of Naples Federico II

Figure 7: one of the 33 racks full of servers

6.1. Big Data

From a decade the Big Data topic is present in every business or industrial function and the interest still solid. Big Data can be defined as a combination of structured, semistructured and unstructured data collected in order to be mined for information. Big Data is often characterized by the so called 3Vs: the large Volume of data in many environments, the broad Variety of data types handled in Big Data systems and the Velocity at which the data is collected and processed. In the big data realm, it is required to process high volumes of low-density unstructured data. The size can range from tens of terabytes of data to hundreds of petabytes. Velocity is the rate at which data is received and processed. Usually, the highest velocity demands are fulfil by a stream directly into memory instead being written to disk.

Traditional data types are structured and therefore suitable for relational database. With the rise of Big Data, data comes in new unstructured or semistructured data types, such as text, audio and photo, which require additional processing to add meaning and support metadata.

Within MAGIC, the Big Data approach will allow a data mining process through all the metadata created, and afterwards through all the images themselves, in order to extract new information. MAGIC will deep drive into the data to extract the key knowledge/Pattern/Information and to create new information from this process, and eventually to make these new information an open access database.

6.2. Internet of Things

Internet of Things (IoT) is defined as an ecosystem of connected physical objects that are accessible through the internet. The “thing” in IoT could be or an automobile with built-in-sensors or a human being monitored by a medical device, or a book with a microchip. In general any device that has the ability to record and transfer data over a network without human intervention contributes to IoT. There are many examples of IoT projects and applications that interconnect people and things, and among the market, the “physical” spread of intelligent objects is also increasing.

The most efficient information repository of the IoT is the RFID (**R**adio-**F**requency **ID**entification) tag: a small two-dimensional microchip (in MAGIC, we will experiment microchips affixed to the original document) which, using electromagnetic induction, prompted by an RFID reader, activates a communication and exchange process to read and write data. Tags (or transponders) are the fundamental unit of the Internet of Things, indeed they allow to activate a much more complete and advanced path for information traceability.

6.3. Artificial Intelligence

Artificial intelligence (AI) generally falls under two broad categories: Narrow AI sometimes referred to as “Weak AI” and Artificial General Intelligence (AGI) sometimes referred to as "Strong AI”.
Narrow AI is often focused on performing a single task and while these algorithms can be defined intelligent, they are operating under strong constraints and limitations. Beside, AGI is a machine with

general intelligence which can apply that intelligence to solve any problem. Narrow AI is, at present time, the most successful realization of artificial intelligence. Few examples are: Google search, Image recognition software and Self-driving cars. Most of Narrow AI is powered by Machine Learning and Deep Learning techniques which uses statistical models to "learn" how to get progressively better at a specific task.

The idea of a "universal algorithm for learning and acting in any environment," (Russel and Norvig) is quite old, but time has not eased the riddle of creating a machine with a full set of cognitive abilities. Nevertheless, the development of the Narrow AI highlights a sustained ferment throughout the world.

Within MAGIC, AI will be used for the image recognition task, which means that within a large set of images, e.g. all the pages of a manuscript or book, AI can reveal if different hands have written it, or can recognize objects within the illustrations of the manuscript. If the analysis is extended to a set of books or manuscripts, AI can help in finding the similarities, if any, in the technique used to write the documents or can help in dating the document itself.

7. Conclusions

This MAGIC project is now just at the beginning, and it is a long term project, but it will allow, in the near future, the digitizazione and fully availability to a large community of the many manuscripts and ancient books, nowadays not accessible but to a very few people chich ara granted the permission to enter the library and handle the books locally. The choice of adopting the same FITS file format used by the Vatican Library, the largest collection of ancient books and manuscripts in the world, will ease the work of scientists who want to make a comparative study of documents stored in different libraries.

8. References

[1] Ambrogio M. Piazzoni 2012 La digitalizzazione nella Biblioteca Apostolica Vaticana *Bollettino di informazione (Associazione dei bibliotecari ecclesiastici italiani)* **3** 7-17
[2] Cesare Pasini 2014 La digitalizzazione dei manoscritti presso la Biblioteca Apostolica Vaticana *Digitalia* **2** 10
[3] Carolin Schreiber, Antonie Magen, Bettina Wagner 2014 New directions and projects for manuscript digitization in German conservation libraries *Bollettino di informazione (Associazione dei bibliotecari ecclesiastici italiani)* **3** 26-32
[4] Carolin Schreiber 2018 Abschlusstagung des DFG-Projekts Erschließung und Digitalisierung von Prachteinbänden als eigenständige Kunstobjekte *Zeitschrift für Bibliothekswesen und Bibliographie* **4** 203-11
[5] Karl-Georg Pfändtner, Carloin Schreiber 2016 Das DFG-Projekt "Erschließung und Digitalisierung von Prachteinbänden als eigenständige Kunstobjekte an der BSB München" – ein Zwischenbericht *Einbandforschung* **39** 6–24
[6] Sunita Barve 2007 *File formats in digital preservation* International Conference on Semantic Web and Digital Libraries 239-48
[7] Donald C. Wells, Eric W. Greisen, R. H. Harten 1981 FITS - A Flexible Image Transport System *Astronomy and Astrophysics Supplement* **44** 363-70
[8] James D. Murray, William vanRyper 1996 *Encyclopedia of Graphics File Formats, 2nd Edition* O'Reilly Media

Acknowledgments
The passion for digital treatment of the ancient manuscripts was born in the University Federico II from a great maestro, prof. Alberto Varvaro, to whose memory this project is dedicated.
The Data Centre has been funded by MIUR, with the project IBISCO (code PIR01_00011), with funds to University Federico II and INFN.