



Virtual Physiological Human Network of Excellence

Grant Agreement: 223920

VPH ToolKit Guideline Document

Topic: Data Characterisation

Version 1.0

20-Mar-11



This page is intentionally blank

Document Information

IST Project Num	FP7 – 2007 - ICT - 223920	Acronym	VPH NoE
Full title	Virtual Physiological Human Network of Excellence		
Project URL	http://www.vph-noe.eu		

Document	Number	G03	Title	Guidance (Data Characterisation)
-----------------	---------------	-----	--------------	----------------------------------

Status	Version.1.0	Final <input checked="" type="checkbox"/>
Dissemination Level	Public <input checked="" type="checkbox"/> Consortium <input type="checkbox"/>	

Authors (Partner)	CNRS	Cervenansky	
	INRIA	Sermesant, Bleuzé	
	UCL	Jacovella	
	UPF	Martelli	
	USFD	Fenner, Mc Cormack	
	UOXF	Cooper	
Responsible Author	Benoît Bleuzé		Email Benoit.Bleuze@inria.fr
	Partner	Sermesant	Phone +33 4 92 38 71 55

Abstract (for dissemination)	This document provides guidance on the attributes required of VPH NoE ToolKit content.
-------------------------------------	--

The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. Its owner is not liable for damages resulting from the use of erroneous or incomplete confidential information.

Version Log			
Issue Date	Version	Author	Change
19-Jul-10	0.1	Benoît Bleuzé	First Draft
31-Jul-10	0.2	WP3	First Issue
24-Nov-10	0.3	WP3-Guidelines taskforce	Re-organisation of the layout and intended purpose
20-Dec-10	0.4	WP3-Guidelines taskforce	Higher level approach.
02-Feb-11	0.5	WP3-Guidelines taskforce	Outline reshuffling
20-Mar-11	1.0	WP3-Guidelines taskforce	Reviewed version, first public release

This page is intentionally blank

Table of Contents

EXECUTIVE SUMMARY	7
INTRODUCTION	8
MOTIVATION	9
<i>General Objectives</i>	10
<i>The Data-Sharing Imperative</i>	10
UNDERLYING CONCEPTS	12
<i>Responsibility</i>	12
<i>Ethics</i>	12
<i>Legislation</i>	12
<i>Provenance</i>	13
<i>Interoperability</i>	13
<i>Sustainability</i>	13
<i>Conclusions</i>	13
INTERACTIONS AND DEPENDENCIES WITH OTHER GUIDELINES	14
<i>Introduction</i>	14
<i>Data Licensing and the VPH</i>	15
<i>Ethics and the VPH</i>	17
VPH DATA CHARACTERISATION.....	19
INTRODUCTION	19
RELEVANT STANDARDS	20
<i>Introduction</i>	20
<i>Ontological Landscape</i>	20
<i>Hierarchical Classification</i>	20
TECHNICAL CHARACTERISATION SPECIFICS.....	22
INTRODUCTION	22
METADATA	24
PROCESSING.....	26
INTEROPERABILITY	27
SUSTAINABILITY.....	28
CASE STUDIES: DATA SHARING PLANS	29
<i>Data-Sharing Plan #1: Biosharing</i>	30
<i>Data-Sharing Plan #2: Cardiac Atlas Project</i>	32
<i>Data-Sharing Plan #3: Medical Research Council Data Sharing Initiative</i>	33
<i>Data-Sharing Plan #4: The Open Provenance Model</i>	34
LEVERAGING THE VPH COMMUNITY	35
<i>Introduction</i>	35
<i>Existing and Planned Activities</i>	35
<i>Dissemination and Training</i>	35
<i>Data Verification</i>	36
CONCLUSIONS	37
REFERENCES	38
APPENDICES	40
<i>Scale-Specific Standards</i>	40

This page is intentionally blank

EXECUTIVE SUMMARY

The Virtual Physiological Human (VPH) Network of Excellence is an EC Community Action that, amongst its many activities, is responsible for establishing an online 'ToolKit' containing software and data that may be of use to VPH practitioners. Contributions to the Toolkit may be made by all members of the VPH community.

This guideline is one of a series of documents intended to assist potential contributors to the VPH NoE Toolkit in preparing their content for submission. The various documents in the series will cover the full range of issues affecting content providers and are being developed over a period of time; when finalised, they will collectively form a single VPH NoE Toolkit resource that may find wider application.

This document examines the challenge of **Data Characterisation**, in the context of data sharing. Raw data is of very limited utility to researchers, as ethical and legal issues, provenance, interoperability and sustainability all introduce constraints on data usage; these key concepts are central to the requirement for well-defined data characterisation, and form an essential set of additional criteria that must be made available by data providers if the VPH is to reap the maximum benefit from the successful sharing of data.

The rich variation in VPH data makes the process of characterisation highly complex. The approach taken in this document is not to create a comprehensive characterisation, but to identify the key issues and examine how these have been resolved in various 'biomedical data-sharing plans' that have evolved in related arenas. Each such standard is then examined from the VPH standpoint and its utility within the VPH is discussed. Many such plans work well in areas of limited scope, and can serve as a basis for the definition of some VPH characteristics; unfortunately no such plan is adequately comprehensive for the very broad domain spanned by the Virtual Physiological Human community. It is hoped that this document will give sufficient insight for researchers to begin publishing data in a meaningful way, and that conversion to any stricter characterisation that may be introduced in future will be straightforward.

Introduction

This document is one of a series that together build to form a complete guide to the ideal content and presentation of materials offered for distribution via the Virtual Physiological Human Network of Excellence ToolKit Portal. The full set of Guideline Documents is summarised below.

Guidance Area	Description
Tool characterisation	The attributes important for inclusion in the documentation of Tools, including performance validation
Model characterisation	The attributes important for inclusion in the documentation of Models, including performance validation
Data characterisation	The attributes important for inclusion in the documentation of Data
Ontological annotation	Methods of knowledge representation, in particular the significance, benefits and methods of ontological annotation of ToolKit content
Interoperability	Key attributes and methods for enabling ToolKit content to be utilised in concert within a multistage workflow
Ethico-legal issues, provenance	The inherited responsibilities that are attached to any item of ToolKit content – perhaps particularly data – including legal, ethical and territorial restrictions
Licensing	The conditions that apply to the legitimate use of the content from an intellectual property standpoint
Usability and training	The factors that are important for the easy use and ready acceptance of ToolKit content, taking into account the environment, the likely users and the need for interoperability. Additionally, the nature of training facilities of all types appropriate to particular content categories.

Table 1: VPH NoE Guidance Documents

Motivation

The VPH ToolKit is designed to support exposure of data and models to a large community. The members of the VPH community have very different goals and working methods, reflecting the wide range of methods and tools in use. This document recognises that exchange of data between practitioners is important, and ideally should be achieved easily through a trusted mechanism. In order to achieve such an objective, these guidelines present a set of imperatives required for data content submission, explaining their impact on interoperability, data integrity, patient confidentiality (or anonymisation given the situation) etc. Moreover, all data should be published with additional information (metadata), that its origin and conditions of acquisition are clear, clarifying provenance so that exchange of data can be undertaken with confidence and trust.

This community already uses many different file formats, and in a wider sense many data types. In order to harmonise workflows, data must be reliably exchanged between tools. This document identifies some of the most common data formats, and introduces a list of criteria, encouraging informed data format choices so that the format best suited to any particular task(s) can be used. In addition to the listing of data types and formats, this document also considers characterisation of data, so that users and developers of tools alike can decide for themselves which data type to use. This includes creation of new formats, guided by a few rules and the application of suitable constraints for a particular usage.

While data has to be generated in a suitable form, there are also advantages in detailing many characteristics of the data (i.e. metadata). Users of images are fortunate enough to have some of these characteristics taken care of in the files' metadata, and metadata is a feature of ontology specifications that sits comfortably with the concept of meaningfully linking disconnected pieces of information. Use cases can illustrate the value of this approach, but they vary significantly. Nevertheless, some advice is presented, along with criteria and priorities that can engender confidence and trust in data published as well as in tools that generate the data.

This document attempts to take into account the range of activity that the VPH encompasses. It tries to cover all sources of data, at different anatomical scales and across scientific disciplines. However, many of the examples are currently imaging related, and this reflects the bias of the authors, who are more familiar with the imaging domain. This will be complemented by increased diversity from the community input, over time.

General Objectives

This document is intended to be concise, to maintain readability. It is a guide designed to help researchers plan their data sharing policy, rather than a mandatory technical framework. It does not represent a strict technical policy, and details will be subordinated to appropriate appendices or linked to more comprehensive related texts. In short the document's objectives are:

- To define the scope of data characterisation principles, namely ethics, law, licensing, reproducibility, interoperability and sustainability.
- To describe the requirements for VPH ToolKit data-sharing based on general quality standards and good practice. This includes:
 - Data technical characterisation
 - Data provenance
 - Data standard formats (ideally including: required metadata, optional metadata, controlled vocabulary to fill in this metadata)
- To introduce examples of existing data-sharing plans and assess their applicability.
- To consider extensions that might form the basis for further work: detailed characterisation criteria to be evaluated as requirements for data submission, a possible taxonomy, and the adoption of ontologies.

The Data-Sharing Imperative

To understand the need for guidance on data-sharing principles it is appropriate first to consider the reasons why data-sharing is desirable and beneficial to research in general and to medicine in particular. An important insight into the benefits of sharing data was given by a significant publication from 1985, aptly titled *Sharing Research Data*[1], which identified several motivations:

- *Reinforcing open scientific inquiry*: ethically Science should be opened to the whole scientific community
- *Verification, refutation, or refinement of original results*
- *Promotion of new data through existing data*
- *Encouraging more appropriate use of empirical data in policy formulation and evaluation*
- *Improvements of measurement and data collection methods*
- *Development of theoretical knowledge and knowledge of analytic technique*
- *Encouragement of multiple perspectives*
- *Provision of resources for training in research*
- *Protection against faulty data*

To this list can now be added a specific imperative that comes from the emergence of the Virtual Physiological Human (VPH) initiative as a significant research activity, namely the

need to make the bulk of anonymised medical data available generally for controlled use in model-based diagnosis and treatment, where access to the widest possible spectrum of information on medical conditions and outcomes is required in order that increasingly personalised care can be built from knowledge of matched data from previous cases.

Finally, the latest imperative comes from the emerging recognition by funding bodies that the effort (and money) put into collecting data for research projects should not be lost after the end of the study. Data should be preserved and made more widely available, and indeed thought concerning the data's ultimate destiny should be given at the study design stage, such that the widest possible subsequent applicability can pertain.

The Medical Research Council [2] reminds the reader on its website that *from 1 January 2006 all funding proposals must include a strategy for data preservation and sharing in the case for support, within a separate section entitled 'Data sharing and preservation strategy'. Any applicants who consider that the data arising from their proposals will not be suitable for sharing must provide clear reasons for not making it available.* The benefits of properly funded data sharing have the potential to far outweigh their cost, hence the message of this section is to encourage readers to plan from the beginning their data sharing policy both technically and financially.

- Time is the first overhead to be included in planning. Searching for the proper formats, filling the necessary metadata fields will take more time than just dumping a data set on an ftp repository without any structure or established provenance.
- Choice of appropriate tools can help in data production and management. Licences may be required, but tools may not even have been developed for the particular case of a specific research topic. In some cases investment in additional effort may be required, e.g. developing tools that will be grafted to a clinical trial protocol to automate the addition of metadata might be essential to the longer-term success of the project.
- Hosting and referencing ones data on servers will generate costs long after the project is over if a local solution is preferred to a publicly hosted solution, which in turn could require a paid subscription.

In conclusion, the world of research data has changed for ever; no longer is data transient and only locally applicable, it now will live on indefinitely. Never has the need for data characterisation been more significant.

Underlying Concepts

Since the early age of medical images, moving data from acquisition equipment to the physician's desk has been a challenge, and similar difficulties apply to researchers and their experimental data. The benefits of exchanging data between research teams can only be harvested if there is sufficient rigour in data publication. Raw data without context is useless to the community, and therefore contextual annotation of data should be considered an essential prerequisite of data submission. This is a process that benefits from application of a few rules, based on some underlying concepts relating to ethics, legislation, responsibility, provenance, interoperability and sustainability.

Responsibility

The VPH is in a position to influence healthcare and for that reason, issues of liability must be given serious attention; consider, for example the issues arising should a VPH interaction adversely affect an individual. There may be many reasons for an unforeseen adverse outcome such as patient variability, databases populated with incorrect data, inappropriate use of data or models, or a misunderstanding of the assumptions associated with a model. The attachment of adequate disclaimers to contributed ToolKit content is perhaps the simplest method of clarifying the terms of use of the resource.¹

Ethics

The presence of an ethical focus is useful in promoting best practice within the VPH whilst avoiding abuse of ToolKit content for direct personal gain. The configuration of the future VPH Portal aims to sustain and support the rights of both contributors and users, encouraging the promotion of positive societal values, and providing an environment that is consistent with the current ethico-legal environment. Respectful consideration of individuals and their protection through adherence to well-defined principles and legislation are key elements, creating a secure environment for resource exchange that does not compromise the ethically driven morays accepted by healthcare professionals. Clearly, significant attention to ethical considerations and ToolKit 'etiquette' is a prerequisite for its success.

Legislation

Since the ToolKit Portal is a repository of contributed health resources (applications, data, knowledge bases, etc.) that may be used to influence patient management (through clinical

¹ As with data protection, the laws (and their interpretation) covering such an eventuality may differ between EU member States and it is difficult to envisage a coherent legal environment in which the VPH can operate. A short-term solution is to establish codes of conduct specific to the VPH, but ideally a longer term solution should be sought, perhaps in the form of harmonization that identifies potential obstacles in each of the member states and proposes effective European solutions. In the meantime, content contributors and users should be aware of their own national obligations in respect of these issues (data protection etc). A list of further information is detailed in the Licensing Guidelines. [STEP report 2006]

decision support, development of new devices by industry, etc.) the principal legislative aspects can be identified as:

- Data protection and freedom of information – this is a mandate in support of human integrity, based around confidentiality concerning the holding and processing of an individual's personal data;
- Liability – responsibility for compensating an individual for injury or mismanagement as a result of interaction with the VPH (regulatory compliance might also be a consideration in some cases);
- Copyright – it is appropriate that all contributions to the ToolKit should consider issues of ownership and clarify the terms of use of the contributed resource. (Licensing is considered elsewhere as a separate ToolKit topic.)
- Licensing – It is increasingly the case that data is made available subject to conditions enshrined in a formal licence agreement.

Provenance

The possibility of tracing data from the processed end result back to the examination or experiment that led to its creation provides a valuable audit trail by which to judge the context and quality of the data. Furthermore, it permits repetition, evaluation of reproducibility, extension to other subjects, refinement of variables and so on, illustrating the benefits of effective metadata annotation. The Open Provenance model[3] is a good example of implementation and definition of the concept.

Interoperability

The value of data is inestimably increased if it can be exploited by peers. However, this is not without legal and technical implications. For example, adoption of a restrictive licence protects those who submit data but hinders its exploitation. Adequate documentation is another aspect that can influence interoperability. Files should be designed to be easily read and written. Documentation of the file format must be available and open to consultation.

Sustainability

Published data is of increased value if it is accessible and available for protracted periods of time. Suitable choice of file format and metadata format can strongly dictate data longevity.

Conclusions

Many aspects of ToolKit content deserve recognition and may indeed require action. This wider perspective imposes additional responsibilities, but these are to be welcomed if an interoperable and sustainable VPH resource is to become a reality.

Interactions and Dependencies with other Guidelines

Introduction

Characterising data is crucial to the distribution of tools that can interact with each other. Such interaction means this Data Guideline document has clear ties with others in the series of guidelines edited by the VPH Network of Excellence. The strongest dependencies are with the two covering Licensing and Ethico-legal issues.

A summary of their relevant content has been added to this document as an introduction to the topics. Understanding these concepts is of prime importance when devising a data sharing policy. These particular subjects are sufficiently large and complex that they deserve their own documentation; however other Guidelines in the NoE ToolKit series also have implications for data characterisation. Their explicit focus may not be data, but readers should be mindful of their existence for the following reasons:

- **Tools Guidelines:** The function of tools is to manipulate data, taking inputs and transforming them into outputs. Characterisation of data and tools necessarily overlaps with descriptions of how files are generated or read, in order to be trusted as sources of useful information.
- **Models Guidelines:** Publication of models involves a protocol and rigor that closely mirrors the exposure (i.e. Publication) of data. Lessons can be learnt from both domains.
- **Ontological Guidelines:** Ontologies have a close association with metadata. They provide a formal representation of knowledge within a domain, listing concepts and the relationships between them. They can be used to reason about the domain and annotated data, automatically making inferences and connections.
- **Interoperability Guidelines:** The guidelines presented under 'interoperability' govern what should appear in a good data definition and documentation. Overlap with this document is necessary because there is no robust interaction if the data is not fully documented and cannot be parsed by other tools.
- **Usability and Training Guidelines:** Users need to be aware of what advantages and drawbacks are inherent to a particular data format. Data documentation should include tutorials or other teaching tools to encourage uptake and interoperability. Details are defined more widely in the Usability and Training Guidelines document.

Since the strongest dependencies of data characterisation relate to Licensing and Ethico-Legal matters, particular justification for their accommodation in data delivered to the ToolKit is given below.

Data Licensing and the VPH

Parallel to the open source movement for software, the concept of open data is emerging. This is part of a wider *open knowledge* field. The term knowledge is taken to include content such as music, films, or books, as well as data be it scientific, historical, geographic or otherwise. Historically there has been relatively little sharing of scientific data. However a wider sharing of data is now becoming common, inspired by high visibility examples such as the Human Genome Project, and an increasing view that the results of publicly funded research, including the data collected or created, should be made openly available. Many funders and journals also now require that raw data is made available.

Openness and licensing for data are crucial to enabling progress in the VPH field, and to scientific progress more generally, by allowing the interoperability of data[4][5]. The volume of scientific data, and the interconnectedness of the systems under study, makes integration of data a necessity. Open data available under a permissive licence is much easier to break-up and recombine, to use and reuse, and hence much more valuable for the common good. The benefits include replication of previous findings, comparisons with independent datasets, testing of additional hypotheses, teaching, and patient safety[6].

The technical challenge of such integration is itself significant, although emerging technologies are helping. But the forest of terms and conditions around data make integration difficult to perform legally in many cases, and the law varies widely between different parts of the world. In many areas, a database may not be openly reusable without an explicit licence granting permission.

Data licensing is thus important because it reduces uncertainty, making explicit whether users can actually use the data, and for what purposes. Supplying this clarity allows researchers to focus on more important matters.

The primary legal framework involved is that of intellectual property, and specifically copyright law. This describes 'works', a kind of property, which have a defined owner (typically the author(s) or their employer, although even this question can be very complex). Copyright law automatically gives the owner of the work certain rights over it, and makes it illegal for others to use the work as though they were its owner; action through the courts can be taken to enforce this. In particular, only the owner can copy, adapt, or distribute the work by default, unless explicit permission is given.

When talking about databases we first need to distinguish between the structure and the contents of a database. Copyright law will generally cover the structure of the database, but may not cover the contents as a collection. Individual contents may be covered depending on their nature. For example, the contents of a database listing the melting points of various substances would not be copyrightable, since they are "facts". Forms of protection for the

contents as a collection fall under copyright law in some jurisdictions, and/or a “sui generis” right for collections of data may exist. Both apply within the EU. Contract law may also be used to protect (closed) databases by providing access only to registered users who have agreed to particular terms and conditions[7].

These variations can create uncertainty or practical difficulty for those wishing to share databases and their underlying data, but retain a limited amount of rights under a “some rights reserved” approach to licensing as outlined in the Science Commons Protocol for Implementing Open Access Data[5].

In order to enhance the utility of data and content shared within the VPH, the VPH-NoE advocates the use of a licence compliant with The Open Knowledge Definition (OKD; <http://www.opendefinition.org/okd/>). Developed by the Open Knowledge Foundation (OKF; <http://okfn.org/>), this outlines the principles that define open knowledge, and by which to judge whether a knowledge licence is open. Within the **Licensing Guideline**, we describe four suitable licences, and categorise each according to 5 key criteria:

- Whether or not it conforms to the Open Knowledge Definition.
- Whether it allows commercial use.
- Whether it is a viral copyleft licence, i.e. it includes a ‘Share Alike’ clause restricting creators of derivative works to the same licence. Such clauses also affect licence compatibility as detailed at the end of this section.
- Attribution – whether you need to acknowledge the work in any derivative of it that you release.
- The legal jurisdiction specified in the licence, if any.

Clinical data, of particular importance to the VPH, poses additional questions that must be considered when looking to release and licence data. Many of the issues are primarily related to ethical considerations, for example the need to de-identify datasets and obtain appropriate ethical approval, ideally with informed consent from participants. These topics are discussed in the accompanying Ethico-legal Guideline. Hrynaszkiewicz et al.[8] also provide practical guidelines and suggest a minimum standard for de-identification. They note that “restrictions on access to certain aspects of data may be warranted, such as when removal of information that could identify the data would negate its scientific value. In circumstances where data must be behind a barrier to universal access, the data could be made accessible only to those who agree to certain conditions of use, and to individuals who meet certain professional criteria. Embargoes on access to data could also be applied.” Such considerations require more specific licensing arrangements than can be covered in these Guidelines, and legal advice should be sought.

Ethics and the VPH

A ToolKit underpinned by sound ethical practice is a recipe for VPH longevity. Ethics mandates respect for the welfare, dignity and rights of all participants interacting with ToolKit content and the VPH. The ethical dimension acknowledges the value of the individual above the processes of biomedical research and requires that such research should be conducted with integrity, honesty, an absence of prejudice, cultural sensitivity, etc. The use of material exploited within the ToolKit for VPH purposes should therefore acknowledge the rights of those who have contributed data and content, including...

- specific rights to publish
- the roles of all contributors
- compliance with relevant legal requirements.

Ethics and Personal Data

With respect to personal data and avoidance of abuse, a legal framework is in place for the protection of EU citizens. This is in the form of the European Data Protection Directive (Directive 95/46/EC), designed to protect the privacy of personal data, and complemented by additional freedom of information measures (e.g. see European Union Directive 2003/98/EC and Regulation (EC) No 1049/2001) which support the right of the individual to inspect the nature of information held about him/her. Further details can be found within the Ethico-Legal guideline document.

Ethics and Health Data

Particular sensitivities are apparent in the context of processing/storing/curating health data since carelessness or abuse² has the potential for significant individual harm³. Ethics is important to the medical profession, which is very careful with personal data and wholeheartedly adheres to long established principles of patient confidentiality. The VPH must adopt similar standards if it seeks any credibility as a biomedical resource. This means that sources of health data contributing to VPH research should have consented to the use of that data, and ideally, should understand the manner in which the data might be exploited (i.e. informed consent). Furthermore, ethical practice requires that such contributors can expect...

2 In 2008, the Lifeblood blood bank reported a data breach involving details of 321,000 blood donors from the loss of two laptop computers. More recently (2009), the similarity of two fax numbers has been attributed to the misdirection of patient details, sent to an Indiana businessman instead of the disability assessment unit of the Tennessee Department of Human Services. The National Health Service in the UK has been credited with over 300 data breaches since Nov 2007 - predominantly the result of theft, but 43 cases involved inadvertent disclosure. These examples are the small tip of a large iceberg. Privacy Rights Clearinghouse estimates that over the 3-year period from 2005 to 2008, in excess of 200 million sensitive personal records were involved in security breaches in the USA alone.

3 Litigation in respect of data breaches is a growing phenomenon. A class action example (United States Court of Appeals) is the case of Ninth Circuit in *Stallenwerk v. Tri-West Healthcare* in which data servers were stolen. The case of Seventh Circuit in *Pisciotta v. Old Nat. Bancorp* (United States Court of Appeals) considered an alleged data breach in which the plaintiffs argued for recompense because the breach resulted in "...substantial potential economic damages and emotional distress and worry that third parties will use [the plaintiffs'] confidential personal information to cause them economic harm, or sell their confidential information to others who will in turn cause them economic harm."

- To be able to withdraw or refuse the use of their data at any time
- That their data is treated with respect (i.e. held for the consented purpose, with adequate measures to protect its integrity)
- That their data is treated confidentially (requires adequate anonymisation procedures, which ensures that inadvertent disclosure of information cannot be linked to an individual)
- That exploitation of their data will not expose them to unnecessary levels of risk.

Principles such as these are explicitly codified in national and European documentation, across Europe and beyond (Directive 2001/20/EC of the European Union; see also *Research Ethics, Committees, Data Protection and Medical Research in European Countries* by Beyleveld D, Wright J, Townend D. Published by Aldershot: Ashgate 2005) but arguably, this is not enough. The reader is reminded that the essence of the ethical mindset is one that asserts the true value of each individual, distinguishing itself by going beyond formulaic procedures, choosing to operate in accordance with nobler principles, such as honesty, integrity, openness, sensitivity, awareness⁴.

Ethics and Submission of ToolKit Content

As a submitter of ToolKit content (noting that the submitted content may consist of contributions from many and varied contributors) it is prudent to consider your action from the perspective of all those involved with the submission of that content and any other people who may be affected by it. Certainly, all persons involved in material content exposed through the ToolKit should have given some consideration to the implications, responsibilities and limitations of data exposure and uptake within the VPH community. How can this be achieved? ... **At the very least read the guideline documents.** These are explicitly designed to inform, and assist researchers with submission of ToolKit content.

⁴ Mahatma Gandhi provides a fitting example of ethics in action in the context of a difficult ethical dilemma. Anecdotes such as this illustrate the tension between ethics norms (codification) and its greater realization in individual cases. "...Early in his career, two of the young people under [Gandhi's] tutelage lapsed into immorality, and [he] agonized for days over a fitting response. Most members of the *ashram* called for strict punishment of the offenders, but it seemed to Gandhi that a guardian or teacher was at least partly responsible for the failure of his ward or pupil. He doubted the other students would realize the depth of his distress and the seriousness of sin unless he did some penance. And so, in response to the students' transgression, he went on a total fast for seven days and took only one meal a day for four-and-a-half months. 'My penance pained everybody,' he concluded, 'but it cleared the atmosphere. Everyone came to realize what a terrible thing it was to be sinful, and the bond that bound me to the boys and girls became stronger and truer'. Cited from Yancey P, *Soul Survivor*, Chapter 7: Mahatma Gandhi: Echoes in a strange land. Published by Hodder & Stoughton 2007

VPH Data Characterisation

Introduction

In the context of the VPH, data characterisation refers to the classification and organisation of data to support interoperability and sustainability. This not only includes the standardisation of data formats, but also data organisation in a flexible framework supporting categories that provide added value to data uses. This is particularly challenging because often categorisation imposes constraints that may be valuable when introduced but become cumbersome and unwieldy as new forms of data emerge. Hence we require the adoption of the concept of extensibility, providing a mechanism to accommodate such a changing landscape, though this can be difficult to implement in practice, as is evident from the modest number of domains in which such principles have been successfully applied.

Relevant Standards

Introduction

The range of standards bodies and standards with applicability to VPH data is so large that it will be the subject of separate and specific examination for a future release of this document. In a future incarnation of the VPH ToolKit it is envisaged that a set of automated features will be provided to assist data-providers in ensuring that an adequate set of standards-based metadata has been provided to make their contributions useful to the community. At the time of writing, there is no such submission server, so these guidelines are aimed at helping researchers to plan submissions according to general concepts. The Metadata discussion in the following section will give a flavour of the desirable information to be included with data files, without undue formalism; this will come as the document evolves with a more comprehensive analysis of standards and with additional contributions from the community.

Ontological Landscape

Central to the work to establish a structured approach to standardisation of terminology, formats and metadata is the work being done by the Ontological community to construct a definitive vocabulary and grammar for the description of VPH terms, and readers are referred to the relevant guideline in this series for further details. Currently there is no single Ontology that encompasses the whole range of domains of interest to the VPH community - we hope the establishment of the following characteristics will help with the prioritisation of the work being led by the Ontology group within the VPH. Some reference materials are included as an appendix to this document.

Hierarchical Classification

Although the provision of a definitive ranking of existing data types is not the aim of this document we can identify a hierarchy of attributes that, in most cases, for most data types, allows an appropriate sorting. Any ranking should at least try to respect the underlying concepts in the following order:

1. Ethics and Legislation
2. Provenance
3. Interoperability
4. Sustainability

The hierarchy of further levels of granularity will depend on case-specific criteria.

The table below illustrates some examples of the widely differing levels of conformance to

standard characteristics of classification, highlighting the reality that at present the need for function often overrides adherence to consistent structure. The classifications used can be debated but this simple example serves to show the difficulties even for popular systems with large user communities, and may provide evidence for the desirability of introducing stand-alone metadata files to help existing formats to cope with the necessities of data sharing:

Data Format	‘DICOM’	‘ANALYZE’
Authorship	Institute and physician tags.	None
Demographics	All patient demographics	Includes a patient ID.
Acquisition settings	Many details encoded, except sometimes in private tags	Limited to some calibration values
Interpretation	Can be a report or link to reports	None
Data Processing	- A DICOM file cannot be modified without issuing A new DICOM file with the corrections mentioned - not every possible processing can be used	None
Interoperability	- Open specifications - So rich and complex no two DICOM readers interpret the format exactly the same way	Open specifications
Sustainability	- Optional additional fields - Very large user base - Community-based adoption of new versions ensure a slow and reasoned update pace.	- Large user base - Superseded by NifTI
Relevance to VPH Researchers	- Quite complex libraries: loading a 3D volume is a complex task	- Historical. Manual segmentations made with Analyze are still revered as reference.

Table 2: Data critique for two popular formats. The traffic-light colouring denotes maturity

Technical Characterisation Specifics

Introduction

This section will discuss the characteristics required of data submitted to the VPH for it to be considered of good quality, and will seek to extend the concepts enumerated earlier. There are several existing characterisations, or formats, that allow the data-provider to embed metadata, and we will take these into account. The reach of the VPH is large and encompasses modelling challenges that in some cases require multi-scale, multi-physics or multi-organ data. This versatility creates the need for interconnected databases, and a unified data-sharing scheme that avoids domain-specific optimisation or omitted parameters.

From the underlying concepts described above, we can construct an initial map of categories:

- **Metadata:** Raw data serves no purpose - the richness of the associated metadata makes the difference between a mute set of bytes and a useful dataset. Note that the term metadata is here used in its broadest sense: all associated information is considered, it need not be embedded in the data file itself to be useful, though it must be accessible. Three types of metadata may be differentiated here: demographical, acquisition and interpretation; in the context of medical data they include:
 - General data, describing the document type and authorship.
 - Demographics, concerning the patient's details, in general anonymised.
 - Acquisition metadata, representing acquisition settings, modality, and geometric information.
 - Interpretation metadata, defining post-study additional data such as diagnosis.
- **Data processing:** Awareness of data loss is a key element here. Conversion and discretisation are common sources of data degradation, and any processing must be described in order to allow repetition of a study, or comparison with other studies.
- **Interoperability:** How easy is it to read or write the data? Documentation of the format and openness of the specifications are the variables.
- **Sustainability:** This characteristic relates to data format, though we should acknowledge the equal importance of the reliability of the physical storage mechanism. Archivists have been studying the issues of digitalisation for some time, and from several prior studies [17], [18] we can draw up a list of sustainability imperatives for file formats:
 - Extensibility (empty optional meta-data fields)

- User base (a large user base assures some longer life cycle)
- Licensing (a very restrictive licence may allow free encoding for several years and then tighten the noose: e.g. H264, JPEG 2000[22])
- Backward compatibility helps in using older files with newer reader/writers.
- Stability (if the format changes every year or so, the confusion about versions is an obstacle to ready diffusion).

Metadata

Well-characterised data requires additional information, 'metadata', to be included along with the raw information. This allows establishing provenance, tracking authors and funding sources, allowing experiments to be reproduced and giving social and demographic data, vital to clinical studies.

Basic Concepts

Data providers should always be mindful of these questions:

- Is metadata embedded in the file (header)?
- Is metadata in a form that is consistent with an external standard?
- Is metadata in human and computer readable form?
- Which parameter(s) does the data describe?
- Which units are used to describe the data?
- What sort of file/data might be submitted?

General Document Descriptive Metadata

For reasons of ownership, responsibility, legal observance, reproducibility and many other factors, all data should identify its authors. There are particular issues of ethics and liability around the identification of authors, and this is discussed in the Ethics guidance document. Following the Dublin Core Metadata Initiative [30], a minimum set of 15 generic metadata fields, allows identification and classification beyond the medical realm – as this allows characterisation of any document, medical or not. The standard increasingly includes extensions to describe more specific domains. The core fifteen items are:

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor
7. Date
8. Type
9. Format
10. Identifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights

Demographics

The second metadata set to be considered is demographics, but field specifics will vary with the scale of the investigation. A key principle is that there should be as much data as possible, to anticipate future work reusing data and taking it in new directions, but perhaps the most vital point is that the data must be properly anonymised. The DICOM standards authority has published a paper on how this can be achieved for imaging data from clinical trials [39].

Acquisition Parameters

Here the added information serves the purpose of enabling replication of the study. To be able to filter data efficiently, to perform powerful searches and eventually build ontologies from the wealth of data to be aggregated we need a **data description**. Additional information includes:

- The **nature of the data**, (in images the modality: CT, MR, temporal and spatial dimensions, etc...)
- The **acquisition equipment**,
- For images, the **geometry** (slice thickness, orientation...).

Scale-specific Data

The remaining categories of technical metadata depend on biological scale and an initial separation of metadata in this way can simplify the work of defining a minimum information set; details will benefit from the specific ontologies for such domains. Currently (2011) initiatives such as the Minimum Information for Biological and Biomedical Investigations (MIBBI, at <http://mibbi.org>) are inadequately advanced to carry data that is equally sophisticated in each field of research.

Processing

Published data is very seldom a raw array of bytes straight from the acquisition apparatus.

- The first processing step is to save the data in a particular format, and a description of that format is required.
- A thorough characterisation will list all processing steps performed on the data, identifying tools, algorithms, and parameters, in sufficient detail for the process to be exactly reproduced.
- Complete reproducibility may be easier when the data is the result of a model, and all parameters are known. Model description languages such as CellML are of significant benefit, and much benefit can be gained from studying their examples [25].

The Open Provenance Model

For a more formal provenance description, submitters are encouraged to examine the ***Open Provenance Model*** (<http://openprovenance.org/>) which, if currently not as well-established as many other open-source initiatives, contains many interesting concepts that could lead to much needed formalisation in the field.

Interpretation Metadata

Any post-processing information can be of value if correctly documented. Interpretation, diagnosis and clinical reporting are of great import to researchers attempting to develop fresh approaches to prognostic data analysis. For imaging purposes, interpretations in 'DICOM report' form [31] are particularly important, as they can be indexed and computationally parsed.

Interoperability

This sub section and the next, sustainability, are predominantly concerned with the file formats used to encode the data, but sustainability is achieved much more easily with interoperability. The relevant characteristics may not yet be prescribed, but should be considered before publishing data, to ensure that the format chosen is adequate and appropriately serves the community. Interoperability is assisted by the use of a format that includes:

- Comprehensive Open Specifications, non-obfuscated or voluntarily obscure.
- Documentation
- Readers/writers/converters, example implementations.

Sustainability

Sustainability parameters are not currently mandatory items when data is submitted to most repositories, but without consideration of relevance to future investigations, the value contained in much data is ultimately significantly diminished.

Important characteristics to consider include:

- Licence (see the licensing section or guidelines for details on restrictive licences)
- Extensibility of the format. Some formats allow for private, optional or future items in their headers.
- The size of the user base, the larger the better
- Backward/forward compatibility
- Degree of maturity of the format, stability

Case Studies: Data Sharing Plans

Several research domains have begun to tackle the issues of data-sharing from a formalised standpoint, and have published their findings. Before creating another such approach for the VPH it is prudent to examine this work and to identify whether the principles and practices adopted have a wider applicability.

By way of example, imaging – a key activity within many VPH activities - has not yet seen much successful effort. However, much better examples are to be found in the Biology and Bio-medical landscapes, and in Genomics, and these success stories offer some inspiration for the VPH.

The following sections contain examinations of several existing Data-Sharing plans, and a future version of this guidance document will examine the applicability of the standardisation included within these and other such plans to a set of exemplar VPH data files, giving an assessment of the level of suitability for adoption.

Data-Sharing Plan #1: Biosharing

Biosharing[9] is the furthest-reaching data-sharing plan identified. Many of its referenced partners or entity members deserve a deeper look beyond the scope of this introductory document, nevertheless a summary of their activity can illuminate how to share data with others when sharing is a key issue.

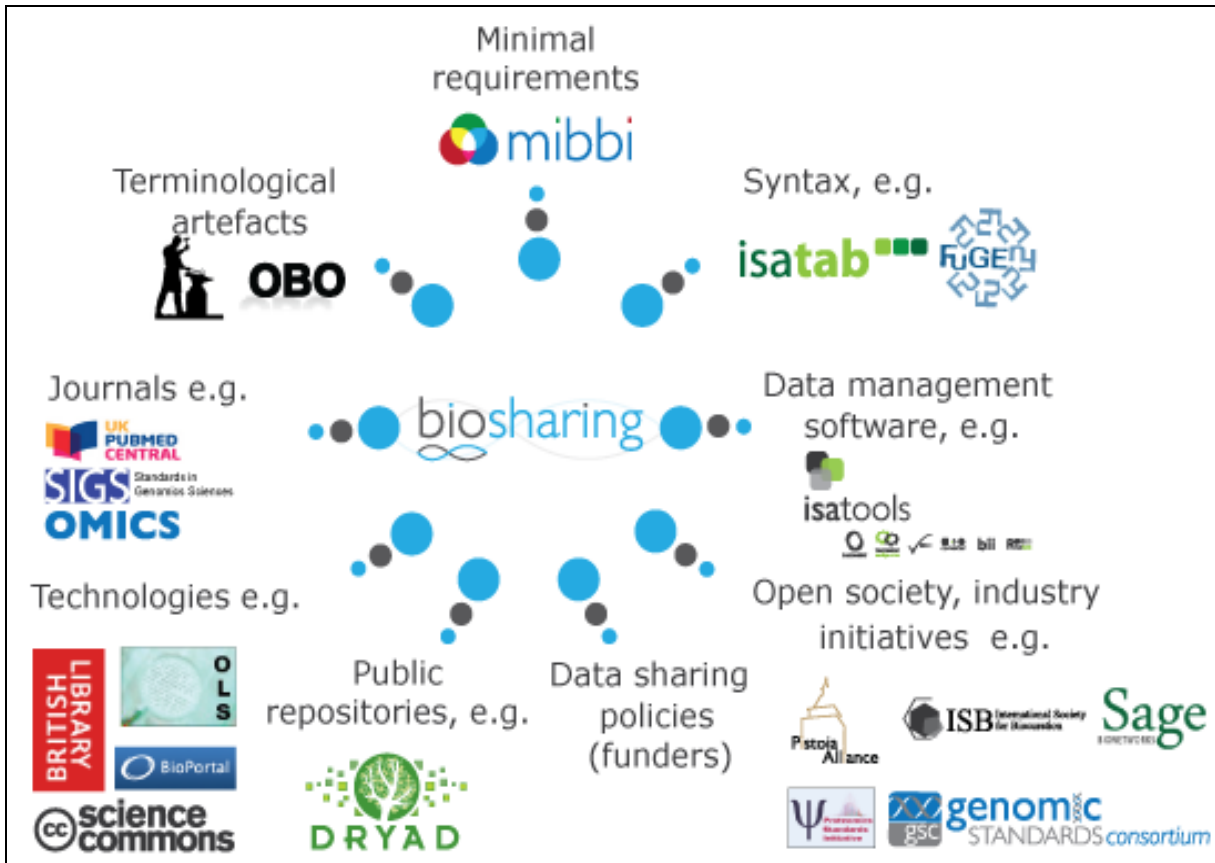


Figure 1: The Biosharing Map.

The number of partners in the Biosharing Map is a measure of its success. Biosharing is an initiative that re-groups journals and repositories of journals, such as OMICS or PubMed Central. The participation of journals is an important factor for the adoption of data sharing policies. As stated in a Nature article: *many journals and funding agencies now require that authors reporting microarray-based transcriptomics experiments comply with the Minimum Information about a Microarray Experiment (MIAME) checklist as a prerequisite for publication*[10]. The list of partners participating in this initiative is long and includes the entire chain of steps necessary in order to publish data. There are several public repositories, including Dryad (<http://datadryad.org/>).

The Biosharing network also links with the MIBBI initiative (Minimum Information for Biological and Biomedical Investigations)[11]. This portal references and creates rules for data sharing policies. Its goals are[12]:

- *To increase the visibility of projects developing guidance for the reporting of aspects of biological and biomedical science.*
- *To encourage collaborative development between such projects, where appropriate, to avoid duplication of effort or competition.*
- *To promote the adoption of consensus guidance on reporting by journals and funders.*

With MIBBI a great number of minimum sets of information necessary to the publication of results in just as many fields of research burgeoned in the past few years. Prior to any submission, we encourage every purveyor of data to have a look at the MIBBI Portal[11]. Many guidelines are included, dedicated to a particular field. For instance one can find there the *Minimum Information about an fMRI Study* (MifMRI) project, a wiki-based checklist. It links to a website with guidelines published on PubMed Central[13].

Ontologies describing the data, allowing links between data also fall under the Biosharing umbrella thanks to the partnership of the Open Biological and Biomedical Ontologies (OBO). They represent a repository of ontologies approved by peers and following a certain set of rules and principles [14].

Data management software has a presence helping with data submission with isatab, which uses ontologies from OBO to generate and help fill out forms.

The Biosharing umbrella extends to members of open societies such as the Genomic Standards Consortium (GSC), whose *mission is to work with the wider community towards:*

- *the implementation of new genomic standards,*
- *methods of capturing and exchanging metadata,*
- *harmonization of metadata collection and analysis efforts across the wider genomics community.* [15]

Conclusion

The Biosharing list is too large to explore in detail, but it demonstrates the potential of cooperative effort when harnessed for data characterisation in pursuit of benefit for the wider community. The harmonisation, categorisation and annotation of data in this way has played a significant part in the rapid deployment of gene sequencing in the last decade.

Data-Sharing Plan #2: Cardiac Atlas Project

The Cardiac Atlas Project [38] gathers DICOM MR images of hearts (normal and diseased) from several contributing institutions. It provides modelling and analysis tools developed at the University of Auckland and is funded by the National Institutes of Health, USA. The goal of this project is to facilitate the development of image analysis, using for instance the statistical analysis of regional heart shape and wall motion characteristics, across population groups, via the application of parametric mathematical modelling tools [38].

The different institutes upload their data through the use of open software developed specifically for the task; this software de-identifies the DICOM images, which can also be annotated with contours and clinical information. Finally the system provides a visualisation facility to display the images and models.

Through a web form, researchers can request access to data, and an interesting particular is in the handling of rights. A potential user of the data will state which data he wishes to use, and the conditions of use of the data, and a committee will review the request on that basis to provide access to some part or the full database. Indeed the participating institutions are funded by different grants and have different motives to share their images. So when contributing data, each contributor describes the rules for using his data. This procedure of reviews seems to be an effective way of encouraging new partnerships with clinicians concerned to maintain control of how their images are used.

Conclusion

The Cardiac Atlas project is an interesting example of a highly-focused, well-controlled strategy for encouraging safe and reliable data interchange.

Data-Sharing Plan #3: Medical Research Council Data Sharing Initiative

The Medical Research Council has put in place a data-sharing policy, and its web site provides a wealth of information:

<http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Datasharinginitiative/index.htm>

The philosophical focus is that of ethical conduct and, although it is not as concrete or comprehensive as the Biosharing initiative, it has the merit of being considerably more pedagogical [15], and is a rich source of information on the key concepts behind data sharing.

Conclusion

The MRC initiative currently offers insights rather than practical facilities, but is perhaps the richest single source of ethical guidance that is applicable to routine data-sharing operations.

Data-Sharing Plan #4: The Open Provenance Model

The Open Provenance initiative is not intended to be a complete data sharing plan, but is concerned more with how to define the provenance of shared data. With the OP Model currently at Version 1.1 [3], it defines an XML format to describe graphical representations of workflows. The Model is described as: *a model of provenance that is designed to meet the following requirements:*

- 1. To allow provenance information to be exchanged between systems, by means of a compatibility layer based on a shared provenance model.*
- 2. To allow developers to build and share tools that operate on such a provenance model.*
- 3. To define provenance in a precise, technology-agnostic manner.*
- 4. To support a digital representation of provenance for any 'thing', whether produced by computer systems or not.*
- 5. To allow multiple levels of description to coexist.*
- 6. To define a core set of rules that identify the valid inferences that can be made on provenance representation.*

Conclusion

Contained in the developing Open Provenance Model may be the seeds of the data-sharing system that will find universal adoption throughout the VPH. It should be required reading for all VPH practitioners.

Leveraging the VPH Community

Introduction

At the time of writing there is no comprehensive automated infrastructure in place within the VPH community to accept and distribute shared data, although this process is being examined within a number of EC-funded initiatives.

Defining a complete taxonomy of categories cannot be done by a single subset of VPH core partners, as the VPH spans too large a domain - this work can only be completed with the help of the entire community. This document, and its implementation in a data server, will evolve with time, thanks to community effort.

Existing and Planned Activities

One of the goals within the VPH initiative is the creation of an extension to the Dublin Core Metadata Initiative that could in time become a metadata standard. The first task is to gather more information and input from future users and submitters that are considered experts in their fields. Gathering a common data description standard across disciplines requires specific input and to obtain this experts are being asked to propose data lists as exemplars for categorisation. Such categorisations are already being attempted in the RICORDO project, a VPH activity that aims to create ontologies from and for VPH communities. Already being planned are separate physiology and anatomical metadata subsets obtained from existing taxonomies such as Mesh[32] and FMA[33].

Dissemination and Training

Consideration of training needs must be given in two separate categories: End Users and Developers.

- The End User – the researcher or clinician who will use tools importing/exporting data in numerous formats - needs to be aware of the characteristics of each data type, the significance of the chosen standard and the wealth of information that may be hidden in the files. Knowing which standard contains what field is a first step towards properly annotated data, and tutorials hosted on the VPH websites, and links to training courses on the data formats community websites, should be considered. At present there are many cases where categorisation is inadequate and workshops disseminating knowledge and train the community are necessary components as the discipline matures.
- The Developer also needs to be trained beyond simply reading documentation, and complex formats such as DICOM are so rich in features that tutorials focus on mere subsets of functionality.

Data Verification

To maintain a consistent database, it is of prime importance to check for data format and metadata conformance with the classification in place. Experience has repeatedly shown hidden complexities in data formats that cause issues of compatibility: DICOM formats include considerable metadata but not a single DICOM reader can open all self-proclaimed DICOM files. Similarly, very few of all the existing DICOM writers (commercial or not) will generate the exact same binary outputs when receiving and encoding the same acquired data, meaning that even without consideration of so-called 'private' fields not present in the Standards, the format is extremely challenging to check for valid implementation.

This exemplifies the need for community activities that go beyond the establishment of theoretical standards to the practical construction of high-quality validation systems for testing both file interpretation and file construction. To that end organisations such as the IHE, Integrating the Healthcare Enterprise[35] *"is an initiative by healthcare professionals and industry to improve the way computer systems in healthcare share information"*. This particular authority tests DICOM compatibility between nodes of all kinds as well as files, and can be understood as a standard body validating DICOM implementations, and is perhaps a pointer to the type of facility that may be required more widely when data-sharing becomes a high-volume reality.

The move towards ontologies is to be welcomed on many fronts, not least because most existing XML based ontology formats come with a schema that allows for initial syntactic metadata file verification, and any infrastructure collecting data has an obligation to perform validation tests. For example, the Biosharing Data-Sharing Plan provides a submission tool 'ISA Tab' that both helps to create metadata and provides automatic validation.

Conclusions

The need for data sharing within the VPH community is significant. The success of such an enterprise depends on well-designed approaches to metadata that provide correct attribution, accurate provenance, and characterised data, allowing browsing, cross-referencing, and providing a longer life to experimental data. This document and the links to its sister guidelines on Legislation, Ethics and Licensing have shed a light on the potential legal hurdles in data publishing, and introduced the key high-level concepts that may assist future data submitters.

The breadth of the VPH community is so large that neither a state of the art analysis nor any document written by a single team can span the extent of data characteristics necessary for comprehensive characterisation. This guidance document has begun the process, establishing a skeleton on which the community can begin to aggregate additional characteristics. The authors cannot overemphasise the need for a full community effort, to extend this guide into a comprehensive data characterisation manual for the benefit of all.

References

1. Sharing Research Data : http://books.nap.edu/openbook.php?record_id=2033&page=R1
2. Medical Research Council Data Sharing policy:
<http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Datasharinginitiative/Policy/index.htm>
3. The Open Provenance Model: <http://openprovenance.org/> and its latest (1.1) specifications : Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E. and Van den Bussche, J. (2010) The Open Provenance Model core specification (v1.1). Future Generation Computer Systems . (In Press)
4. Open Knowledge Foundation on openness and Licensing:
<http://blog.okfn.org/2009/02/02/open-data-openness-and-licensing/>
5. <http://sciencecommons.org/projects/publishing/open-access-data-protocol/>
6. <http://www.bmj.com/content/340/bmj.c181.full>
7. <http://www.opendefinition.org/guide/data/>
8. Hrynaszkiewicz et al. (2010) : <http://dx.doi.org/10.1136/bmj.c181>
9. BioSharing : <http://biosharing.org>
10. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project :
<http://www.nature.com/nbt/journal/v26/n8/pdf/nbt.1411.pdf>
11. MIBBI Portal: http://mibbi.org/index.php/MIBBI_portal
12. Goals of MIBBI: http://mibbi.org/index.php/About_us
13. Guidelines for Reporting an fMRI study (listed from MifMRI):
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2287206/?tool=pubmed>
14. The OBO foundry principles: http://obofoundry.org/wiki/index.php/OBO_Foundry_Principles
15. The GSC mission: http://gensc.org/gc_wiki/index.php/GSC_Mission
16. Medical Research Council Data Sharing Initiative:
<http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Datasharinginitiative/index.htm>
17. **Digital Formats: Factors for Sustainability, Functionality, and Quality**. Paper by Caroline R. Arms and Carl Fleischhauer for IS&T Archiving 2005 Conference, Washington, D.C. (http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml)
18. Digital Preservation Guidance Note 1: Selecting file formats for long-term preservation. Adrian Brown, Head of Digital Preservation Research, UK National Archives, August 2008
<http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf>
19. Some data sharing policies referenced on the Biosharing website, being from European or American institutions: <http://otter.oerc.ox.ac.uk/biosharing/?q=policies>
20. DICOM (**D**igital **I**maging and **C**ommunications in **M**edicine) : <http://medical.nema.org/>
21. JPEG Committee drafts: <http://www.jpeg.org/jpeg2000/CDs15444.html>
22. Patent issues in JPEG 2000. Here is the point of view of the web master on the JPEG website:
http://www.jpeg.org/faq.phtml?action=show_answer&question_id=q3f042a5e42fd8

23. Cuellar, A.A., Lloyd, C.M., Nielsen, P.F., Bullivant, D.P., Nickerson, D.P. and Hunter, P.J. "An overview of CellML 1.1, a biological model description language" SIMULATION: Transactions of The Society for Modeling and Simulation International. 2003 Dec;79(12):740-747.
24. Lloyd CM, Halstead MD, Nielsen PF. "CellML: its future, present and past" Prog Biophys Mol Biol. 2004 Jun-Jul;85(2-3):433-50.
25. CellML Meta data specifications: <http://www.cellml.org/specifications/metadata>
26. HL7 (Health Level 7 International): <http://www.hl7.org/>
27. JCAMP-DX committee: <http://www.jcamp-dx.org/>
28. Gene Ontology: <http://www.geneontology.org/>
29. MINC file format reference:
http://en.wikibooks.org/wiki/MINC/Reference/MINC2.0_File_Format_Reference
30. Dublin Core Initiative: <http://dublincore.org/>
31. <http://www.pixelmed.com/srbook.html>
32. <http://www.nlm.nih.gov/mesh/>
33. <http://sig.biostr.washington.edu/projects/fm/>
34. RICORDO, *project focused on the study and design of a multiscale ontological framework in support of the Virtual Physiological Human community to improve the interoperability amongst its Data and Modelling resources* <http://www.ricordo.eu/>
35. IHE (Integrating the Health care Enterprise):
36. <http://www.ihe.net/Connectathon/index.cfm>
37. A lot of work has already been done by the IHE organisation on validating DICOM files. Their Connectathon may be the largest test bed for DICOM file conformance:
<http://www.ihe.net/Connectathon/index.cfm>
38. Cardiac Atlas Project: <http://www.cardiacatlas.org>
39. DICOM Supplement 142: Clinical trial De-identification:
ftp://medical.nema.org/medical/dicom/supps/sup142_14.pdf

Appendices

Scale-Specific Standards

Molecular Level

- JCAMP-DX[27], Bruker formats for spectrometry data. Many information cannot be included in DICOM format.
- **Gene Ontology**[28] (GO) focuses on genes and proteins.

Organ Level

- **Digital Imaging and Communications in Medicine (DICOM)** is the de facto standard for medical images acquisition and exchange. Its copyrights belong to the **National Electrical Manufacturers Association (NEMA)**, and it is being developed by the DICOM Standards Committee. DICOM defines two concepts: a network transfer protocol as well as a file format. The transfer protocol concerns the interoperability guidelines more than the data characterisation guidelines, thus we will from now on concentrate on the file format.
- JCAMP[27], Bruker formats for spectrometry data.
- MINC[29] from NetCDF format. Allow to store more meta-data than DICOM for some specific disciplines.
- The **Foundational Model of Anatomy (FMA)** is an ontology of human anatomy.

Patient Level

- **Health Level 7 International**[26] (HL7) is mostly not a standardisation body for data, but for data exchange. HL7 mainly defines messaging and interaction protocols. They however developed the **Clinical Document Architecture (CDA)** which “*is an XML-based markup standard intended to specify the encoding, structure and semantics of clinical documents for exchange*”. It is a container for documents.