



ACCÈS UNIFIÉ AUX DONNÉES
ET DOCUMENTS NUMÉRIQUES
DES SCIENCES HUMAINES ET SOCIALES

CONTRACT N°K1432

Hosting of IT services and data for Human and Social Sciences in France

Date:	31 January 2008
Version:	1.0
Document status:	Final
Author	Olof Barring

List of acronyms	3	
Introduction	4	
Executive summary	4	
Service constituents.....	5	
Application provider.....	6	
Service provider	7	
Hosting Environment Provider	7	
Database Administrator	8	
Long Term Preservation provider	8	
Data centre survey.....	9	
General questions about Scientific and Technical Information (STI) management services	10	
Data volume and service load	10	
Long Term Preservation	11	
Service details	12	
Service and support level.....	13	
Computer centre resources	14	
Survey of existing STI service providers for the Human and Social Science community in France		15
Relevance of Grid technologies.....	17	
End - user survey.....	19	
Appendix A: data centre survey.....	20	
General questions about Scientific and Technical Information (STI) management services	20	
Service details	23	
Service and support level.....	27	
Computer center resources	29	
Appendix B: Questionnaire pour les fournisseurs d'IST pour la communauté SHS en France		33
Appendix C: End user Survey.....	44	
References.....	49	

Version	Date	Editor	Comments
0.1	09.01.08	Olof Barring	Initial version
0.2	11.01.08	Olof Barring	Added observations to the STI service provider survey
0.3	17.01.08	Olof Barring	Implement language corrections from Tony Cass and comments from Benoît Habert
0.4	24.01.08	Olof Barring	Add TELMA answers to DC survey. Add section on Grid relevance. Add prototype end-user survey
0.5	25.01.08	Olof Barring	Implement comments from Benoît Habert
0.6	29.01.08	Olof Barring	Add logos and final comments from Benoît Habert
1.0	31.01.08	Olof Barring	Final document

List of acronyms

DBA	Database Administrator
HA	High Availability
LTP	Long Term Preservation
OAI	Open Archive Initiative
OAIS	Open Archive Information System
OGF	Open Grid Forum
OS	Operating System
RAID	Redundant Array of Independent Disks
SHS	Social and Human Sciences ¹
STI	Scientific and Technical Information
UPS	Uninterruptible Power Supply
WAN	Wide Area Network

¹ In English the acronym SSH 'Social Sciences and Humanities' is frequently used but here SHS is kept in line with the French 'Sciences Humaines et Sociales'

Introduction

The objective for this study is to review from an IT perspective the existing Scientific and Technical Information (STI) and Long Term Preservation (LTP) services provided for the Human and Social Sciences in France and to propose possible rationalizations, for instance in terms of sharing of IT resources. The scope does not cover a comparative evaluation of the existing services nor any other qualitative aspects viewed from a Human and Social science perspective.

Most information presented in this report is based on two questionnaires:

- A detailed Data Centre survey of existing or potential future hosting and LTP environments for SHS, and
- An STI service provider survey with the objective of understanding the service requirements on the hosting and LTP environment.

Executive summary

The set of sites and STI service providers reviewed in this study is by no means claimed to be complete. This is partly due to the limited time and partly to the heterogeneity and size of the Human and Social Science community in France, which made it difficult to obtain a full list of service providers and hosting centres from the beginning. Another difficulty was that two centres, which at the same time act as hosting sites and as service providers, declined to participate in the survey:

- INIST did not give any particular reason
- CRDO Aix-en-Provence gave as reason that they are waiting for clear statement concerning their status and future from TGE/Adonis

It should thus be kept in mind when reading the conclusions and recommendations below that this is based on a sub-sample of hosting sites and service providers. Another aspect that is not covered by this study is the hosting of services and data for sociology, economy or health science communities, which may have specific access control and privacy requirements.

Having analysed the answers to the two questionnaires and visited the three largest data centres (CINES, CC-IN2P3 and JOUVE) covered by the survey, I make the following recommendations. These recommendations are independent and can therefore be implemented in any order.

Recommendation 1 Concentrate LTP to a (few) large Data Centre(s) as a centralised service provided to all STI services for SHS in France. A logical choice from the IT perspective would be building upon the existing expertise at CINES and TELMA OAIS starting with file formats currently supported. In terms of IT infrastructure the service is best hosted at CINES or CC-IN2P3, where the latter alternative would require moving (or sharing) the expertise. For an appropriate funding the additional costs for the increased resources should be evaluated by the selected site(s) once the requirements (e.g. in terms of data volume and supported formats) are known. Addendum: Regularly mirror an incremental copy of the data archive for SHS from CINES to CC-IN2P3 or vice-versa for disaster recovery reasons and also fail-over in case the backup site has its own LTP expertise.

Recommendation 2 Ensure the provision of High Availability, load-balance and/or hot-standby solutions for the hosting of STI services and associated components (e.g. the database) for service applications that support this. Work with providers of other applications to ensure support for such solutions.

Recommendation 3 Allocate funding to CC-IN2P3 for a second person for the integration and operation of new STI services for the SHS

Recommendation 4 Require documented procedures for routine operation and service recovery as part of the production deployment of an STI service for the SHS.

Recommendation 5 Minimise the operating system types and versions required for the hosting environment. This study is not concerned with the actual choice of a common OS base but *Linux* seems to be commonly supported by many of the providers. Application providers should be asked for an estimate of the feasibility and cost of the porting and packaging of their software to Linux (or whatever the selected OS may be).

Recommendation 6 Centralise the hosting of STI services for SHS in a small number of large data centres, either commercial or government funded (where CINES and CC-IN2P3 are strong candidates). For each STI service the application and service providers should assess their hosting requirements for the proposed environment and if possible adapt the application/service. The hosting provider should provide training (how to use the hosting environment), support and some flexibility allowing for a progressive handover of low-level system administration and operation tasks.

Recommendation 7 Undertake, with involvement of the end-user communities, an independent assessment of all Digital Library type applications to establish if there are sufficient differences to motivate maintaining different applications and software development teams. If/Where appropriate, launch (a) software development project(s) to design and implement (a) common service application(s).

Recommendation 8 Require the (re)design of STI services such that neither root (administrator) access for the service manager to the production server nodes nor database administrator privileges are required during normal running. The hosting provider should propose a development environment where the STI service manager can develop and test new features/fixes.

Recommendation 9 Require applications to automatically re-establish database connections in the event of an intermittent loss due to a High Availability switch-over of the service.

Service constituents

The SHS community in France is a heterogeneous collection of independent scientific user communities with rather different requirements for STI and LTP services. In order to distinguish common requirements from the specific needs among the different scientific communities it is useful to define a set of service constituents that can easily be mapped into IT roles and skills. From discussions with existing service providers for the SHS communities as well as some scientific users the following constituents have been identified:

1. Provider and maintainer of a scientific application that covers, in full or in part, a particular STI service

2. Service provider of an STI service
3. Provider of the hosting infrastructure for one or more STI services
4. Database Administration
5. Provider of the LTP service for the data produced or required by an STI service

The five service constituents and the corresponding customer-provider relationships are depicted schematically in Figure 1 and will be described further in the following subsections.

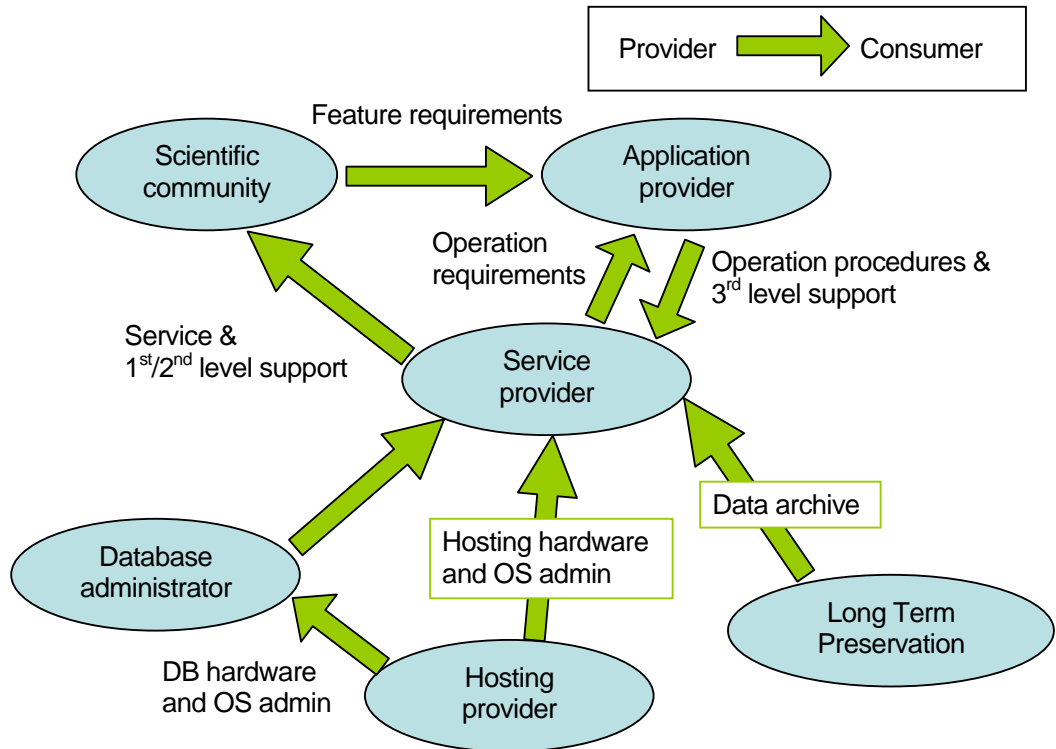


Figure 1: customer provider relations

Application provider

An application is a piece of software developed to fulfil some scientific purpose. Three examples are:

- A (web-based) dissemination interface to query and display some scientific data or document stored in a LTP archive. If the archive is OAIS compliant such dissemination interface implements a *Finding Aid*.
- Generation or processing of scientific data. The output data will either require LTP or serve as input for further processing by other applications leading eventually to end results which require LTP.
- Collaborative platforms enabling scientists to work together remotely on common projects, where the output presumably requires LTP. Example: the ARCHEOVISION Virtual Reality centre (Odéon) in Bordeaux

Providing software to a scientific community requires software engineering skills and usually also a good understanding of the scientific community for whom the application is provided.

Service provider

A service provider is responsible for the management of the STI services, which includes the deployment and operation of the applications that together form the service. "Second level" (expert) user support is usually also part of the service management. In the absence of a helpdesk or a call centre, service management also includes the first level user support.

It is often difficult to draw a clear line between the application provider and service provider as the latter usually needs deep insight of the functioning of the application and has usually also operational requirements on the application itself. It is not unusual for both roles to be fulfilled by the team developing the application.

Equally it is often difficult to draw a clear line between the service manager and the provider of the hosting environment. It is, however, in this case necessary to clearly define the responsibilities as it is rarely motivated that both roles are fulfilled by the same team because the skills required are different. Typical issues that must be clarified are:

- Login access, both privileged (root/administrator) and unprivileged, to the server machines
- Scheduling of intrusive interventions (e.g. kernel or hardware upgrades)
- Application performance monitoring (probing)

See references [1], [2] and [3] for a set of best practice and de-facto standards governing Service Management and System Administration (see next section).

Service management requires IT engineering skills and usually also good scientific insight.

Hosting Environment Provider

The provider of the hosting environment is responsible for low-level system administration as well as for issues such as high level capacity planning, procurement and the network and electrical setup.

Examples of low-level system administration tasks requiring IT technician skills are:

- Operating System (OS) installation
- Deployment of regular updates released by OS vendor
- Housekeeping, e.g. cleanup of the */tmp* partition
- Administration of storage infrastructure, e.g. RAID configuration, hard disk replacement

Examples of high-level administrative tasks requiring IT engineering skills are:

- Capacity planning and procurement

- Design of high-availability deployment of an STI service together with the application and service providers
- Host and infrastructure security
- System configuration, performance (e.g. CPU utilization and load average) and exception monitoring
- LAN administration and where relevant fibre channel fabric configuration

Database Administrator

The Database Administrator (DBA) is responsible for the deployment, tuning and operation of the databases that a service depends on. Depending on the complexity and size of the database system used, the role may be fulfilled by the service manager. Certification courses exist for complex systems like Oracle. The Data Base administrator role requires IT engineering skills and is usually provided by the hosting environment provider.

Long Term Preservation provider

The provider of LTP need not be geographically or even administratively close to the hosting environment. LTP concerns not only long-term storage of the actual data onto physical media but also preservation of its *understandability*. LTP providers should comply with the concepts described in the Open Archive Information System (OAIS) reference model [4] and derived standards. In particular the responsibilities listed in section 3.1 of the OAIS reference model and repeated below for convenience must be undertaken by the organization providing Long-Term preservation:

- *Negotiate for and accept appropriate information from information Producers.*
- *Obtain sufficient control of the information provided to the level needed to ensure Long-Term Preservation.*
- *Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided.*
- *Ensure that the information to be preserved is **Independently Understandable** to the Designated Community. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information.*
- *Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original.*
- *Make the preserved information available to the Designated Community.*

As mentioned previously, an important duty for the LTP provider is to maintain the *understandability* of the data. In practical terms this implies a responsibility to understand the original data format and to follow its evolution over time. If the data format specification undergoes a non-backward compatible change, it is the responsibility of the Long-Term preservation provider to perform a reformatting (logical migration) of the original data to conform to the new specification. The success of a reformatting operation depends on the conformance of the original data and, for instance, use of non-standard extensions may cause the reformatting operation to fail.

This means that the LTP provider must *verify* that the format conforms to the specification already at data ingestion. For widespread formats like HTML, PDF, PDF/A, JPEG, PNG this is facilitated by the fact that the formats are defined in an open and published specification and for some of them (e.g. PDF) there are even tools to automatically verify the conformity. See [5] for a relatively detailed list of various file formats and links to their specifications (where available).

The LTP provider is also responsible for the physical media on which the data is stored. The data should be copied to new media (physical migration) before the current hardware gets obsolete or there is a risk for deterioration. Large data centres with tape infrastructure regularly perform complete physical migrations of the accumulated data every 3-5 years.

Data centre survey

In order to understand how STI services are hosted for the SHS in France today a questionnaire targeting salient features of the IT infrastructure and staffing was developed. This questionnaire was divided into four main sections:

- 1) General aspects about the service: target scientific community, information type (text, image ...), methods used for client authentication and authorization management, current number of users and resource utilization and expected trends for the next two years.
- 2) Specific details of the STI services and data preservation: setup and warranty details of the hardware used for hosting the STI services and data storage; service quality aspects like use of high-availability (HA) solutions, redundant storage solutions (e.g. RAID) and backup, network and power redundancy.
- 3) User (scientist) support arrangements and coverage
- 4) General IT infrastructure: total data centre CPU and storage capacity, tape (or other mountable media) infrastructure, power and cooling, operation staffing and coverage

The following centres hosting STI services for the SHS in France were invited to participate in a survey

Short name	Data centre		Hosted STI service(s) for SHS	completed survey
	Full name			
CINES	Centre Informatique National de l'Enseignement Supérieur		http://www.persee.fr/ , http://www.abes.fr/	yes
CNRTL/ATILF	Centre National de Ressources Textuelles et Lexicales / Analyse et Traitement Informatique de la Langue Française		http://www.cnrtl.fr/	yes
CRDO Aix-en-Provence	Centre de Ressources pour la Description de l'ORAL / Laboratoire Parole et Langage		http://crdo.up.univ-aix.fr	no
CRDO Paris/LACITO	Centre de Ressources pour la Description de l'ORAL / Laboratoire de Langues et Civilisations à Tradition Orale		http://crdo.vjf.cnrs.fr	yes
CRN/M2ISA	Centre de Ressources Numériques Méthodologies pour la Modélisation de l'Information Spatiale Appliquée		http://www.m2isa.fr/	partial
CC-IN2P3	Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules		http://ccsd.cnrs.fr/ , http://www.cn2sv.fr/	yes
INIST	Institut de l'Information Scientifique et Technique		http://biblioshs.inist.fr/	no
JOUVE				yes
TELMA/ ENC	Traitement ELectronique des Manuscrits et des Archives / Ecole National des Chartes		http://www.cn-telma.fr/	Yes

The centres that answered the survey fall into two categories:

- Small to medium size data centres, 'Centre de Resource Numérique' (CRN), where the STI service and data storage for SHS is the only major service hosted by the centre: CNRTL, CRDO Paris, and M2ISA
- Medium to large size data centres where the STI service and data storage for SHS is one service among others hosted by the centre: TELMA/ENC², CINES, CC-IN2P3 and JOUVE

M2ISA services and archive are hosted at the 'Fondation de la Maison des Sciences de l'Homme de Paris' (FMSH). As this was created only recently the archiving service will only be deployed during 2008 and thus they could only partially complete the questionnaire. M2ISA was therefore invited to also complete the STI service provider questionnaire in the next section.

Apart from 'Ecole National des Chartes' (ENC) TELMA also hosts 'Institut de recherche et d'histoire des textes' (IRHT).

JOUVE is a commercial company (<http://www.jouve.fr>), which already hosts some STI services and data for SHS.

The detailed answers to the questionnaire are given in Appendix A. The sections that follow give the salient conclusions drawn from the answers to the questionnaire and from observations during visits to the three data centres CINES, CC-IN2P3 and JOUVE.

General questions about Scientific and Technical Information (STI) management services

Data volume and service load

The Human and Social Science data is moderate in volume between 1 – 20TB today. The data scattered over a large number of objects with an average size ~1MB, except for CNRS/LACITO where the average object size is 273MB. The data volume is expected to at least double in the coming 2 years with some variation between the sites where the largest (CINES) expects 50TB, which is still moderate even in today measures.

Summary of the data ingest from information provider users:

- The number of information provider users is usually relatively small, less than 50 for the participating sites and expected to grow to at most 300 in coming 2 years. This is an advantage given that adding or updating data in general requires a strict access control, for which there is an administrative overhead (e.g. account expiry)
- As can be expected from the moderate total data volume the average inbound data rate generated by the information providers is low: 5MB/s for CNRS/LACITO and 3MB/s for JOUVE. For the other sites, who did not provide an answer, it is assumed to be low

Summary of the data access from information consumer users (e.g. scientists):

² TELMA is classified in this category because they provide >10 other services and STI services for SHS are not the main ones

- The number of unique users accessing the stored information varies significantly between the sites. CINES is largest with 300,000 users today and an expected ~50% growth in coming 2 years.
- Such large number of users is not a problem when the information is publicly accessible (OAI [6], Open Archives Initiative) but the administrative overhead of user management could become a burden in other cases.
- The daily number of access and outbound data rate does not necessarily depend on the total volume of data or number of objects.
- For CNRTL/ATILF the average rate is almost 10 accesses per second aggregating to an outbound data rate of 400KB/s, with an expected 50% growth in coming 2 years. CNRTL/ATILF did not provide a number for the current use space but the allocated disk space is 300GB. Depending on how the service application is designed this can be considered as a rather loaded service for a relatively modest data volume.
 - It could be well motivated with a load-balanced front-end cluster.
 - The disk subsystem has to serve a constant random access rate, which motivates an enterprise class subsystem, which is also what CNRTL/ATILF is using.

Long Term Preservation

As explained previously the LTP goes far beyond the mere archiving of data on some physical media. A sustainable LTP service requires:

- Personnel investments to build up the sufficient expertise for development and operation of an automation framework for the data ingest, archive and access in compliance with the OAIS reference model. The personnel skills include IT engineers for software development and operation but also an archivist verifying the process
- Storage infrastructure and operation that prove to scale with the anticipated accumulated data volume. Storage media (disk or tape) include mechanical and electromagnetic parts, which require stable environmental conditions for safe operation. For instance, spinning disks should be operated with
 - redundancy to disk failures (RAID)
 - UPS (Uninterruptible Power Supply) and safe power (diesel generators) coverage
 - sufficient cooling to minimise risk for media deterioration
- Nevertheless the media wears out or becomes obsolete and part of the storage operation includes periodically procuring new storage hardware and undertake migrations of the total accumulated data volume to new the media

Following those requirements, it is recommended to centralise the LTP to a (few) large Data Centres providing the service for all STI services.

Among the data centres participating in the survey, only CINES and TELMA appear to have undertaken a serious effort to adopt the OAIS reference model and the derived standards for LTP. As for both CINES and TELMA this has been a significant

investment (e.g. 5-6FTEs of IT engineering efforts for CINES) it is recommended to build upon this service beyond the local use.

Currently CINES can provide this service for some text and image formats. The cost of the service depends strongly on the diversity of formats and also data volume. As will be discussed later CINES also benefits from 3x2Gbit WAN links through the Renater network and has therefore in principle no network limitation for serving as a central LTP for data from other sites.

Independent of this report there has been a suggestion that CINES should focus on its core business, High Performance Computing (HPC), and the LTP activity to be moved elsewhere, e.g. to ABES. This can still be in line with the above recommendations provided that the selected Data Centre is sufficiently large and fulfils the criteria listed above.

The TELMA OAIS compliant LTP service supports XML based formats (XML/TEI, XML/EAD) and some digitized image data format. One important limitation, however, is the relatively small size of the Data Centre in terms of staffing and infrastructure. For instance, the tape backend only consists of a single manually operated LTO-2 drive and the Data Centre does not have any dedicated cooling capacity.

The experience from both CINES and TELMA has shown that it is not only requiring a substantial effort to develop an OAIS compliant LTP service but also to maintain the expertise and administer the production service. This experience motivates a concentration to a (few) large site(s) as a centralised LTP service that is open for use of STI services for SHS in France.

Recommendation 1 Concentrate LTP to a (few) large Data Centre(s) as a centralised service provided to all STI services for SHS in France. A logical choice from the IT perspective would be building upon the existing expertise at CINES and TELMA OAIS starting with file formats currently supported. In terms of IT infrastructure the service is best hosted at CINES or CC-IN2P3, where the latter alternative would require moving (or sharing) the expertise. For an appropriate funding the additional costs for the increased resources should be evaluated by the selected site(s) once the requirements (e.g. in terms of data volume and supported formats) are known. Addendum: Regularly mirror an incremental copy of the data archive for SHS from CINES to CC-IN2P3 or vice-versa for disaster recovery reasons and also fail-over in case the backup site has its own LTP expertise.

The addendum suggesting a mirrored copy is discussed later.

The lack of OAIS compliant Long Term Preservation was discussed during the visit to JOUVE. This has not, until now, been required by any of their current customers nor is there any other principal business motivation to provide such a service. However, JOUVE is very flexible in terms of tailoring solutions for their customers and OAIS compliance could be part of the offer. It should be noted, though, that, as the expertise has to be built up and an automation framework has to be developed, there is likely to be a cost impact, at least for the first customer requesting such a service.

Service details

CRDO Paris/LACITO did not answer any question in this section. It is therefore assumed that *no* High Availability, data protection or redundancy setup is used. M2ISA only partially answered this section, which is probably because the data archive is not yet in production.

Data integrity and availability considerations:

- all centres except CRDO Paris/LACITO perform regular incremental or full backups
- JOUVE, CINES, CC-IN2P3, CNRTL and TELMA/ENC all use some RAID option, which together with the backups minimises the risks of data loss and, as the RAID disks also are hot-swappable, maximises the data availability

STI service availability considerations:

- JOUVE, CINES and CC-IN2P3 all provide both network and power redundancy.
- CNRTL and TELMA/ENC have power (UPS) redundancy only
- JOUVE and CC-IN2P3 provide High Availability options for their STI services while CINES only use it for their LDAP service.
- Use of High Availability (HA) solutions may not necessarily be required by the Human and Social Science user community but since presumably service downtimes must be announced and scheduled there are operational benefits in being able to do maintenance work without general service interruption.
- As the number of machines required for the STI services is anyway relatively low, it appears as an unnecessary risk if an unrecoverable failure of a single machine on a Friday evening results in a complete service outage for the whole weekend. For sites with 24/7 operational coverage a configured hot-standby machine together with an operator procedure would be sufficient.

Recommendation 2 Ensure the provision of High Availability, load-balance and/or hot-standby solutions for the hosting of STI services and associated components (e.g. the database) for service applications that support this. Work with providers of other applications to ensure support for such solutions.

Service and support level

The answers to the questions in this section refer to two different types of ‘users’ depending on whether the centre is merely responsible for the hosting or also for the STI service management. In the first case, technical support is provided to the STI service manager while in the second it is provided to the scientific user community (see Figure 1). JOUVE, CINES and CC-IN2P3 provide the hosting and archive (and LTP in the case of CINES) but not the STI service. CRDO Paris/LACITO, CNRTL/ATILF and TELMA/ENC provide both. As noted previously, CRN/M2ISA currently only provides the STI service with the hosting provided by ‘Fondation de la Maison des Sciences de l’Homme de Paris’ (FMSH).

In the model where the provider of the hosting environment is not the same as the STI service provider, it is important to clearly define the responsibilities.³ Another difficulty in this model is that the initial deployment of the STI service in the hosting environment usually requires a non-negligible effort from the hosting provider to train the STI service and application providers in how to use the hosting environment. The application provider usually also needs to make software and packaging changes to meet the operational requirements of the hosting provider. In the last couple of years CC-IN2P3 has acquired extensive experience of hosting foreign services from the

³ Further discussed later in this report when considering the conclusions drawn from the responses to the second survey ‘*Questionnaire pour les fournisseurs d’IST pour la communauté SHS en France*’.

Medical and Biological sciences and is now applying the same model for SHS: c.f. the recent success hosting 'Centre National pour la Numérisation de Sources Visuelles' (CN2SV). Although the initial effort for the hosting provider can be substantial this will decrease with time as the STI service provider learns how to use the new environment. CC-IN2P3 relies on a single person for this activity, which is not a sustainable situation if the hosting of foreign services is to be extended further.

Recommendation 3 Allocate funding to CC-IN2P3 for a second person for the integration and operation of new STI services for the SHS

STI services can benefit from outside working hours support at data centres having some level of 24x7 operational coverage (e.g. JOUVE, CINES and CC-IN2P3) provided that documented procedures for common service problems are available. Care should be taken when possible to design the procedures for technician level staff. CINES has formalized this as part of their service level agreement, requiring the foreign STI service manager to provide a hardcopy of the procedures for routine service operational tasks and complete service recovery.

Recommendation 4 Require documented procedures for routine operation and service recovery as part of the production deployment of an STI service for the SHS.

Computer centre resources

A reduction in the set of hardware and OS types required for the STI services is desirable as proliferation inevitably leads to higher hosting costs. Of the services running in the data centres participating in this review, the majority seem to be supported on some Linux flavour. One important exception, however, is ABES which requires Windows platform with a SUN based storage backend. As the ABES cluster is the only main Windows cluster operated at CINES, the development cost of porting ABES to Linux should be compared to the ongoing costs for maintaining the Windows administration expertise dedicated for ABES at CINES.

Recommendation 5 Minimise the operating system types and versions required for the hosting environment. This study is not concerned with the actual choice of a common OS base but *Linux* seems to be commonly supported by many of the providers. Application providers should be asked for an estimate of the feasibility and cost of the porting and packaging of their software to Linux (or whatever the selected OS may be).

As one of the Tier-1 centres for the high energy physics experiments at the Large Hadron Collider (LHC), CC-IN2P3 has a tape storage capacity significantly greater than any of the other centres participating in the survey. Furthermore, in comparison to the requirements of the LHC experiments, the aggregate data storage required for the LTP of the data from SHS is marginal. As an addendum to the **Recommendation 1** above, it is therefore suggested that CINES regularly (daily) exports to CC-IN2P3 an incremental copy of its data archive for Human and Social Sciences for disaster recovery purposes.

Although the answers to the questions of the survey do not reveal any particular technical problem with hosting infrastructure for the smaller centres where the STI service for the Human and Social Science is the main service, it is unlikely to be cost

effective in the longer term to maintain small dedicated IT infrastructures and their associated staffing. It is instructive, for instance, to compare the personnel cost relative to the total budget at CC-IN2P3 (20%) to that at the smaller centres (>90%). However, as it is likely that at the smaller sites the same staff is responsible for both the service management and the hosting, one cannot expect an immediate gain in manpower by centralizing the hosting since the service management responsibilities remain. Nevertheless, such a centralisation would be beneficial for the end users since it allows the smaller centres' IT engineers to focus on service management and support while the system administration and operation tasks are performed by dedicated IT technician level staff in the hosting Data Centre. In terms of candidates for centralised hosting, CC-IN2P3 and CINES must be considered as viable options, but there are alternatives. Commercial companies like JOUVE, 'Internet Fr' (<http://www.internet-fr.net/>) and others should be considered, but preferably on the basis of a bulk offer (in response to a call for tender) for hosting the various STI services already operated today for the Human and Social Sciences.

Recommendation 6 Centralise the hosting of STI services for SHS in a small number of large data centres, either commercial or government funded (where CINES and CC-IN2P3 are strong candidates). For each STI service the application and service providers should assess their hosting requirements for the proposed environment and if possible adapt the application/service. The hosting provider should provide training (how to use the hosting environment), support and some flexibility allowing for a progressive handover of low-level system administration and operation tasks.

The **Recommendation 1** and **Recommendation 6** are independent. The centralised hosting of the STI services may well be different from that of the LTP as long as the wide area network (WAN) bandwidth is sufficient. Both CC-IN2P3 and CINES are connected to Renater. JOUVE has currently 3Mbits/s WAN connectivity but is upgrading to 100Mbit/s, which is not high but probably sufficient for SHS.

Survey of existing STI service providers for the Human and Social Science community in France

The second questionnaire developed for this study concerns the providers of the Scientific and Technical Information (STI) services. It specifically targeted service providers who do not provide the hosting. The aim of this questionnaire was twofold. Firstly, to clarify with real examples how the sometimes complicated organization of responsibilities between the service- and the hosting- providers has been arranged. Secondly, establish which features of the hosting environment are perceived as the most important for the service providers. In order to facilitate and attract the service providers to complete this questionnaire it was translated to French. The questionnaire and detailed answers are given in Appendix B.

The five STI service providers participating in the survey each use a different hosting environment provider, which makes the survey particularly interesting as it brings out their differences. Some of the Digital Library type services (EDP, Persée and Revues.org) may cover more than just Human and Social Science.

All service providers develop and maintain the service application, at least within the same organization. Whilst this is certainly convenient there should be a serious assessment as to whether or not there is sufficient commonality to justify a shared code base (and associated development team), at least for the Digital Library type services. The assessment should consider:

- Savings in man-power efforts by having a single large software development team rather than many small
- If not the whole application whether or not at least some parts of it could be jointly developed and maintained, e.g. client authentication/authorisation layer, database interface layer, search engines, publish/subscribe techniques, graphical user interface
- Advantages for the hosting provider if the services can be deployed and operated following a common set of procedures
- Advantages for the end user SHS scientists that may have to consult multiple Digital Libraries and would benefit from a homogenous look-and-feel and query interfaces

Although they did not take part in this study, it appears from their web-site (<http://www.inist.fr>) that INIST (Institut de l'Information Scientifique et Technique) is working in this direction.

Recommendation 7 Undertake, with involvement of the end-user communities, an independent assessment of all Digital Library type applications to establish if there are sufficient differences to motivate maintaining different applications and software development teams. If/Where appropriate, launch (a) software development project(s) to design and implement (a) common service application(s).

The involvement of the end-user (scientist) community is vital since the success of moving to a common service application depends entirely on whether or not it meets the aggregated needs from the different user communities.

LTP: all five STI services store some sort of text format and could therefore presumably use the OAIS compliant LTP services provided at CINES following **Recommendation 1** above, which is the case already for the largest provider in terms of storage volume Persée.

For security reasons, the service manager should not need root (administrator) access to the machines. Any operation requiring root access should be delegated to the hosting provider staff. Enabling remote interactive login (non administrator) is also a security risk but may be inevitable due to the complexity of the service management tasks, which include investigation of runtime problems. It is interesting to note that the 5 providers all require different login access to the server nodes, as this hints that it is a difficult problem and the hosting providers have to be flexible. The limited or no login access adopted by EDP and CN2SV (Centre National pour la Numérisation de Sources Visuelles) is doubtless the preferred model.

The hosting provider should in parallel propose a 'development' environment where (if the site policy allows for it) a more privileged access could be allowed for the STI service provider to develop and test out new features and fixes before being deployed in production.

Recommendation 8 Require the (re)design of STI services such that neither root (administrator) access for the service manager to the production server nodes nor database administrator privileges are required during normal running. The hosting provider should propose a development environment where the STI service manager can develop and test new features/fixes.

The separation of the hosting and service responsibilities is in general good although it may be desirable that the low-level system monitoring and housekeeping tasks (e.g. cleaning temporary directories) are performed by the hosting provider staff.

Concerning the database systems used, there is a nice clustering around MySQL and PostgreSQL, both of which are well established and usually well known to hosting providers. The separation of the Data Base Administrator (DBA) responsibilities is good. **Recommendation 6**, that suggests that the hosting is centralized to the large Data Centres, implies that the hosting provider is also responsible for the main DBA routine tasks (apart from the optimization of the application itself). In particular, low-level tasks such as the placement of the database files should be performed by the hosting provider staff.

Use of High Availability is desirable as it allows the DBA to perform transparently interventions like patch upgrades etc, which would otherwise need to be scheduled together with the service provider. However, deployment of an HA solution alone is unlikely to be beneficial; application software must be capable of automatically re-establishing the database session.

Recommendation 9 Require applications to automatically re-establish database connections in the event of an intermittent loss due to a High Availability switch-over of the service.

As stated in the **Recommendation 8** above, the application should not require database administrator privileges during normal running. The argument is the same as for root (administrator) access to the servers. Application vulnerabilities to SQL injection exploits could have disastrous effects to the database if the application runs with administrator privileges.

In general it seems that the STI service providers are satisfied with their hosting environment with one exception, CRN/M2ISA, who are hosting their services and archive at 'Fondation de la Maison des Sciences de l'Homme de Paris' (FMSH). Unfortunately they do not answer to the follow-on question about what have been the main problems with their provider but rate that machine stability and availability as well as the quality of the system administration are very important for them. Otherwise, the high satisfaction of the STI service providers with their hosting environment shows that it is beneficial also for STI services for SHS to separate the development and management of a service from the hosting of it. This supports the **Recommendation 6** suggesting centralization of the hosting providers.

As the centralization requires moving the service and its data, there are a number of considerations out of which the STI service providers rate as most important the moving of the data and maintaining access to the service while it is moved. The integration of the STI service in the new hosting requirement is not considered an important problem.

Relevance of Grid technologies

Grid enabled federations of Digital Libraries and Digital Archives have gained large research and development interest both in the US and EU. An Open Grid Forum (OGF) Research Group called "Preservation Environments" (PE-RG) [7] is looking at demonstrating LTP concepts on federated Data Grids. Such an environment could potentially be an alternative to a centralised LTP as suggested in **Recommendation 1** but it should be emphasized that this is a Research Group and apart from a requirements document, it has not published much progress. Centralised solutions

based on production ready technologies (e.g. from CINES and/or TELMA) allow for rapid deployment of a shared service addressing the immediate LTP needs in SHS. In the medium term the progress of the research around Grid enabled federations of Digital Libraries and Archives should be reviewed to understand if the technology has reached a level of automation where it is no longer necessary to maintain centralised centres of expertise.

The mirroring of the LTP archive to CC-IN2P3 suggested for disaster recovery and failover reasons in the addendum to **Recommendation 1** could very well be implemented with Data Grid tools for replication management, for instance the Logical File Catalogue (LFC) and File Transfer Service (FTS) that are part of the EGEE [8] gLite software stack.

The EU project "DILLIGENT" [9], which was partially funded by the Sixth Framework Program, had as objective to create a test-bed built on integrated Grid and Digital Library technologies for allowing e-Science organisations to access shared knowledge and facilitate collaborations. The project started in September 2004 with 36 months duration. The DILLIGENT consortium did not include any major French partner.

A follow-on project "D4SCIENCE" has been proposed in the Seventh Framework Program [10].

Another Digital Library related project proposed in the Seventh Framework Program is "DRIVER" [11], where CNRS is one of the partners and contributes with the HAL [12] archive system maintained by CCSD and hosted at CC-IN2P3.

CLARIN is a pan-European initiative with the ambitious goal of establishing an integrated and interoperable research infrastructure of language resources. To achieve this it combines technologies and ideas from Data Grid, Digital Libraries and Semantic Web. The French partners in CLARIN are CNTRL/ALTIF and TELMA/DIS which have both participated in the Data Centre survey of this study, and ELDA (Evaluations and Language resources Distribution Agency) in Paris.

In the Open Grid Forum (OGF) there is now a so called Community Group (CG) for Humanities, Art, and Social Sciences (HASS-CG) [14]. The group does not seem to be very active and it does not yet have a charter.

ARCHEOGRID [15] is a project launched at the University of Bordeaux 3 for archiving of 3-dimensional modelling data and associated scientific documentation of visually reconstructed partially (or fully) destroyed archaeological sites. It can be considered as a National Grid Initiative (NGI) for SHS in France and though the goals look similar it does not seem to be linked with an Italian (INFN Florence) project with the same name.

The 'Institute Des Grilles' [16] [17] is a CNRS unit formed in 2007 for coordinating French NGI activities. For 2008 a major exercise is planned to review the needs of various scientific communities, including SHS.

In summary there are a number of punctual grid-oriented initiatives for SHS in France but so far a global strategy for the community as a whole seems to be lacking. Firstly it is likely that the SHS community is large and the needs may be different among the sub-communities. Another reason, however, may simply be lack of awareness or at least a clear view of how to make use of the shared computing resources made available through grids. This situation should improve through the initiatives launched by 'Institute Des Grilles' but it should be kept in mind that enabling a scientific application to successfully and efficiently run on a grid usually requires a non-negligible effort. The motivation for taking this step may naturally arise when the local computing resource becomes the limitation. Centralisation of LTP and hosting large

data centres, as proposed in **Recommendation 1** and **Recommendation 6** , with dedicated efforts for helping SHS user communities integrating their applications in a foreign run environment may speed up this process.

End - user survey

An end-user survey was not part of the scope for this study. Such survey could however be motivated for two reasons:

1. Increase the SHS user awareness of existing STI services and how to use them
2. A comparative study of the end-user satisfaction of the existing STI services could pinpoint features or issues that should be considered when deploying similar new services for other SHS communities

The second point is also motivated for a successful implementation of **Recommendation 7** above. A prototype survey given in Appendix C is based on input from a few SHS scientists in France.

Appendix A: data centre survey

General questions about Scientific and Technical Information (STI) management services

The purpose of these questions is to understand if your data center is already providing STI services for SHS in France and, if so, for which communities

Does your data centre already provide IT services for Human and Social Sciences in France?		
Company:	No	Yes. Please give details of user community(ies)
JOUVE		www.annee-philologique.com (old literature) www.memoiredeshommes.sga.defense.gouv.fr (Differents wars) www.masson.fr (books and revues) for INIIST (credit card payment) www.societedesetudeslatines.com
CINES		Persée portal (http://www.persee.fr/) Liberfloridus (http://liberfloridus.cines.fr/) SUDOC catalog (http://www.sudoc.abes.fr/) Canal-U CEFAEL (http://cefael.efa.gr/) Couperin (http://www.couperin.org/)
CRN/M2ISA		archéologues, géographes, historiens, anthropologues, ethnologues, linguistes
IN2P3		Hébergement de sites de données de recherche en SHS pour le Centre National pour la Numérisation de Sources Visuelles Support et infrastructure pour le CCSD pour la plate-forme HAL (notamment HAL-SHS et le consortium NorBib des SHS en Scandinavie) Hébergement de sites Web Hébergement des données et de la plate-forme revues.org (via le CCSD) Projet en discussion pour la fourniture de puissance de calcul et stockage pour le projet ArcheoVision
CRDO		storing, conservation and access
Paris/LACITO		
CNRTL/ATILF		Principalement pour les communautés des linguistes et littéraires, mais aussi pour l'ensemble des chercheurs ayant besoin d'exploiter des corpus, dictionnaires et lexiques
TELMA/ENC		Ces services s'adressent prioritairement aux chercheurs et unités de recherche en histoire médiévale et moderne, en lexicographie ; ensuite, aux institutions de conservation qui ont la charge des documents primaires mis en ligne ; enfin, aux utilisateurs finals qui incluent ces communautés et aussi des étudiants, des amateurs d'histoire, etc.

Does your data center provide information archive and preservation services to other communities?		
Company:	No	Yes. Please give examples of user community(ies)
JOUVE		- bank (without details, i think you know why) - results about invitation to tender -
CINES		Long-term preservation of : 1. PHD theses from ABES, 2. digitized documents from Persée
CRN/M2ISA	No	
IN2P3		Physique des particules: données et calcul des expériences de physique des particules, physique nucléaire, astro-particules et astrophysique Biologie et recherche biomédicale: stockage de données et calcul
CRDO		throw oai and web portals
Paris/LACITO		
CNRTL/ATILF		Oui, c'est le cas pour des collègues linguistes, nous confiant leurs ressources, à terme nous devrions offrir ces mêmes service dans le cadre de la MSH Lorraine et du réseau des MSH
TELMA/ENC	No	

For which type(s) of information does your data centre provide archive and preservation services?							
	<i>Text</i>	<i>Images</i>	<i>Audio</i>	<i>Video</i>	<i>Local bibliography databases</i>	<i>Mixed documents (manuscripts with images + text, XML collection files, ...)</i>	<i>Other (please specify)</i>
<i>Company:</i>							
JOUVE	X	X		X		X	
CINES	X	X	X	X	x	X	
CRN/M2ISA	x	X	X			X	- Bases de données spatiales (géographiques) structurées selon les normes géographiques internationales de l'ISO / TC 211 : données spatiales dynamiques au format d'échange international shp, grid (raster dynamique), au format de transport e00, - GeoDatabases - Données Spatiales numériques dynamiques en 2D et 3D dans des formats d'échange conformes aux normes géographiques standards de l'ISO / TC 211 accompagnées de leurs métadonnées selon la norme géographique internationale iso 19 115 / TC 211 - Orthophotoplans, satellitaires au format Geo-TIFF - photographies aériennes au format TIFF ou Geo-TIFF root, dicom, fits, base de donnée objet, 3D
IN2P3	X	X	X		X	X	
CRDO Paris/LACITO	X	X	X	X			
CNRTL/ATILF	X						Corpus, dictionnaires, lexiques, outils informatiques de traitement de la langue
TELMA/ENC					x	x	

Are there any proprietary or privacy aspects of the information that require authentication and authorization of providers and consumers?				
	<i>None (all information is public)</i>	<i>Legislation or information owner may require you to provide client identity</i>	<i>Confidentiality may require client authorization, i.e. authenticated clients are subjected to access control depending on which information is accessed.</i>	<i>Other (please specify)</i>
<i>Company:</i>				
JOUVE				X
CINES			X	
CRN/M2ISA	x		X	
IN2P3				X
CRDO Paris/LACITO				X
CNRTL/ATILF	x			
TELMA/ENC	x			
				Données biomédicales confidentielles et personnelles anonymisées
				Il est actuellement possible pour l'utilisateur final de créer un compte (nom et mot de passe), afin d'annoter les corpus, les annotations étant privées.

User authentication method?			
Company:	X509 grid certificates	kerberos	Other (please specify)
JOUVE	X		Login/mot de passé
CINES			login + password
CRN/M2ISA			
IN2P3	X	X	user/mot de passé
CRDO Paris/LACITO			apache authentication
CNRTL/ATILF			
TELMA/ENC			Compte utilisateur géré par un SGBDR (MySQL)

How are the authentication and authorization managed?			
Company:	LDAP (Lightweight Directory Access Protocol) based, e.g. Active Directory, IBM Tivoli Directory Server, Fedora Directory Server	NIS (Network Information Service)	Other (please specify)
JOUVE	X		database Radius
CINES	X		
CRN/M2ISA	X		
IN2P3	X		
CRDO Paris/LACITO			flat file
CNRTL/ATILF			
TELMA/ENC			Compte utilisateur géré par un SGBDR (MySQL)

Human and Social Science information provider customers (e.g. Journal Editors): what is the current and projected utilisation of information provider customers (document, images, ...)?									
Company:	Used storage space (GB), current	Used storage space (GB), projected for next 2 years	Number of stored objects (documents, images, ...)	Number of unique users, current	Number of unique users, projected for next 2 years	Number of upload requests per day, current	Number of upload requests per day, projected for next 2 years	Average inbound data rate (KB/s), current	Average inbound data rate (KB/s), projected for next 2 years
JOUVE	1000	1100	>1000000	11000	visists by day			3 mbs	
CINES	30000	50000	10000000	5	10	10	500		
CRN/M2ISA	1000-5000	20000			100		1000		
IN2P3	500	2000	463000						
CRDO Paris/LACITO	958	3000	3500	30-50	100-300	10	20	5120 KB/S	5120 KB/S
CNRTL/ATILF									
ENC	2	15		7	25				

Human and Social Science information consumer customers (e.g. Human or Social Scientists accessing items): what is the current and projected utilisation of information consumer customers (document, images, ...)? The input boxes accept free format text so you may give ranges (e.g. 5 - 10)

<i>Company:</i>	<i>Number of unique users, current</i>	<i>Number of unique users, projected for next 2 years</i>	<i>Number of access (e.g. download) requests per day, current</i>	<i>Number of access (e.g. download) requests per day, projected for next 2 years</i>	<i>Average aggregate outbound data rate (KB/s), current</i>	<i>Average aggregate outbound data rate (KB/s), projected for next 2 years</i>
JOUVE						
CINES	300 000	400 000	36 000	40 000		
CRN/M2ISA		1000		10000		
IN2P3						
CRDO	70	100	3000	10000		
Paris/LACITO						
CNRTL/ATILF			500000	750000	400	600
TELMA/ENC	522		750		51200	

Please provide any other relevant details, e.g. compliance with the OAIS (Open Archival Information System) reference model, of the Scientific and Technical Information management services provided by your data centre?

<i>Company:</i>	
JOUVE	we can realize software development, ergonommy, interface design, referencement
CINES	100% OAIS compliant usage of ISAG(G), ISAAR(CPF) recommendations DMCI metadata implementation + other french standards (P2A, PSE)
CRN/M2ISA	Prévisions : - Accès à des géoportails nationaux et internationaux, - Accès aux réseaux de données géographiques (Geography Network) - Géotraitements en temps réel des données spatiales et/ou thématiques implantées sur un autre site distant MSH(travail collaboratif)
IN2P3	
CRDO	
Paris/LACITO	
CNRTL/ATILF	
TELMA/ENC	Compatibilité de la plate-forme avec les principes du modèle OAIS.

Service details

Detailed questions about the IT services provided for management of Scientific and Technical Information (STI)

What amount and type of storage is allocated for the Scientific and Technical Information management services at your site? please specify volume in usable GigaBytes (raw capacity minus overhead from filesystem, RAID, etc.).

<i>Company:</i>	<i>Online (disk)</i>	<i>Nearline (automated libraries with tape, optical disks or other mountable media)</i>	<i>Offline (shelved tapes, optical disks or similar media requiring manual mounting)</i>
JOUVE	1000		
CINES	32000		
CRN/M2ISA			
IN2P3	5000	0	0
CRDO			
Paris/LACITO			
CNRTL/ATILF	300		
TELMA/ENC	3		3

Is the Scientific and Technical Information management services data backed up on independent storage, either a second copy on a different disk system or a copy on tape?

<i>Company:</i>	<i>No Independent dual copy is created when information is uploaded</i>	<i>Regular incremental backups</i>	<i>Regular full backups</i>	<i>Other (please specify)</i>
JOUVE		x	x	but not independent, all customers are mixed
CINES	x	x		
CRN/M2ISA		x	x	
IN2P3		x		
CRDO				
Paris/LACITO				
CNRTL/ATILF		x		
TELMA/ENC			x	

Data integrity control. All storage media deteriorate with time and data risks to be corrupted. A dual copy (e.g. backup) reduces the risk for data loss. What other measures are taken for assuring the integrity of the data stored in with the Scientific and Technical Information management services? For instance, is the checksum (md5 or other) calculated at upload time and monitored at every access?

<i>Company:</i>	<i>None</i>	<i>Checksum calculated at upload time and verified at every access</i>	<i>Checksum calculated at upload time and regularly verified even if the data is not accessed</i>	<i>Other (please specify)</i>
JOUVE		x		we use checksum process but not for our human science customers
CINES			x	
CRN/M2ISA				
IN2P3				Le centre de calcul dispose de plusieurs système de stockage. Certains d'entre eux proposent des commandes d'upload et de download avec checksum. Aucun de nos système effectue un checksum sur les données non accédées. Cette opération est impossible sur notre centre étant donnée la volumétrie que nous gérons. L'utilisateur peut implémenter son propre checksum sur ces données.
CRDO				
Paris/LACITO				
CNRTL/ATILF	x			
TELMA/ENC	x			

Hardware hosting the service application. The purpose of this question is to understand how many machines are used for running the Scientific and Technical Information management services today and the growth trend over the last couple of years.

<i>Company:</i>	<i>Total number of machines hosting the STI services today</i>	<i>Total number of CPUs/cores hosting the STI services today</i>	<i>Total number of CPU/cores hosting the STI services end of 2005?</i>	<i>Number of CPU/cores added in 2006?</i>	<i>Number of CPU/cores retired in 2006?</i>	<i>Percentage of machines under warranty (0 = none, 100=all)</i>
JOUVE	17	23	13	4	0	100
CINES	30	50				100
CRN/M2ISA		en cours de création, opérationnel en mai 2008				
IN2P3	5	34				100
CRDO						
Paris/LACITO						
CNRTL/ATILF	3	6				100
TELMA/ENC	1	2	2	2	2	100

Warranty details. When a hardware failure occurs, please specify the intervention conditions required of the vendor(s). Please make the distinction between working hours and other hours (e.g. nights and Weekends). If you have different warranty contracts depending on machine type, please specify a range, e.g. '4-8' or give multiple answers, e.g. '4,8'.

<i>Company:</i>	<i>Maximum time to intervene (hours, 0 = not specified)</i>	<i>Maximum time to intervene (working hours, 0 = not specified)</i>	<i>Maximum time to problem solved or hardware replaced (hours, 0 = not specified)</i>	<i>Maximum time to problem solved or hardware replaced (working hours, 0 = not specified)</i>	<i>Other (please specify)</i>
JOUVE	4	2	0	4	
CINES	0	4	0	8	
CRN/M2ISA					
IN2P3			day+2		
CRDO					
Paris/LACITO					
CNRTL/ATILF	0	4	0	8	
TELMA/ENC		24		24	

If more than one machine is used for hosting the Scientific and Technical Information management services, do you use load-balancing or other high availability (HA) solutions?

<i>Company:</i>	<i>No load-balancing or other type of HA is used</i>	<i>Fixed mapping of users over the different servers</i>	<i>DNS load-balancing</i>	<i>HA-linux</i>	<i>Vendor specific HA solution</i>	<i>Other HA solution (please specify)</i>
JOUVE				x	x	tomcat's connector (mod_jk)
CINES	x					Linux HA Heartbeat for LDAP services
CRN/M2ISA						
IN2P3						piranha
CRDO						
Paris/LACITO						
CNRTL/ATILF	x					
TELMA/ENC						une seule machine pour le STI

Network redundancy: if more than one machine is used to host the Scientific and Technical Information management service, what is the redundancy against failures of the local area network devices?

<i>Company:</i>	<i>No network redundancy: machines are on the same switch</i>	<i>Redundancy against switch failures (machines on different switches)</i>	<i>Redundancy against router failures (e.g. switches connected to multiple routers)</i>	<i>Other (please specify)</i>
JOUVE			x	x
CINES			x	
CRN/M2ISA				
IN2P3				x
CRDO				
Paris/LACITO				
CNRTL/ATILF	x			
TELMA/ENC				une seule machine pour le STI

Electric power redundancy: how are the Scientific and Technical Information servers (and all other necessary devices and services e.g. network routers) protected against power cuts?				
<i>Company:</i>	<i>No redundancy against main power failure</i>	<i>Machines are connected to uninterruptible power supplies (UPS)</i>	<i>Machines connected to UPS and site secure power supply (e.g. Diesel Generators) protecting against long periods of power outage.</i>	<i>Further details (e.g. UPS coverage)</i>
JOUVE		x		x 2 differents power supplies by host
CINES CRN/M2ISA IN2P3 CRDO Paris/LACITO CNRTL/ATILF TELMA/ENC		x		x

Characteristics of the disk subsystem used for the Scientific and Technical Information management service. One can usually classify disk subsystems as either Desktop or Enterprise type. The former targets low price per GigaByte while the latter offers high levels of data availability. Please note that subsequent questions below will allow you to further qualify the disk subsystems.			
<i>Company:</i>	<i>Desktop class disk subsystem</i>	<i>Enterprise class disk subsystem</i>	<i>Further details (e.g. main vendor)</i>
JOUVE			HP
CINES CRN/M2ISA IN2P3 CRDO Paris/LACITO CNRTL/ATILF TELMA/ENC		x	Marque : HP Vendeur : Alphamega

How is the disk hardware connected to the service hosts? Hot-swappable means that the disk can be changed without service interruption.						
<i>Company:</i>	<i>Disk trays integrated in host chassis. The disks are not hot-swappable</i>	<i>Disk trays integrated in host chassis. Disk are hot-swappable</i>	<i>External disk arrays with single host attach</i>	<i>External disk arrays with dual host attach</i>	<i>Hosts and disk subsystems on storage area network (SAN)</i>	<i>Other (please specify)</i>
JOUVE		x		x		
CINES CRN/M2ISA IN2P3 CRDO Paris/LACITO CNRTL/ATILF TELMA/ENC		x				X X X

Disk redundancy configuration. What RAID (Redundant Array of Independent Disk, http://en.wikipedia.org/wiki/Redundant_array_of_independent_disks) configuration is used? Depending on the disk subsystem referred to in the previous questions, the vendor would normally propose hardware RAID controllers. Some vendors also propose solutions with integrated software RAID (e.g. <http://www.sun.com/servers/x64/x4500/specs.xml>). The most common RAID options are: * RAID-0 : disk striping (not for redundancy) * RAID-1 : disk mirroring * RAID-5 : disk striping with parity disk * RAID-6 : disk striping with two parity disks

Company:	none	hw RAID0	hw RAID1	hw RAID5	hw RAID6	sw RAID0	sw RAID1	Sw RAID5	sw RAID6	Other (please specify)
JOUVE			x	X						
CINES				X	x					
CRN/M2ISA										
IN2P3				X				X		
CRDO										
Paris/LACITO										
CNRTL/ATILF				X						
TELMA/ENC				X						

Service and support level

Details about the service and support level agreement with the customers to the Scientific and Technical Information (STI) management service

Do you have a formal written agreement with your customers for service and support levels for the Scientific and Technical Information management services provided at your data center?

Company:	No	Yes	Formal service level agreement, SLA. (please give a brief description or link)
JOUVE		x	SLA
CINES		x	Standard SLA for information providers
CRN/M2ISA		x	Charte des utilisateurs
IN2P3	X		Au moment de la demande de compte, les utilisateurs signent un document dans lequel ils certifient avoir lu la charte de bon usage de nos ressources informatiques. Il est possible d'avoir un accord écrit concernant le niveau de sécurité et fiabilité nécessaire pour le projet.
CRDO			
Paris/LACITO			
CNRTL/ATILF	X		
TELMA/ENC		X	

If applicable and the information can be disclosed, please indicate how you charge for the service

Company:	Not applicable (or cannot be disclosed)	No charging	Only information provider is charged (e.g. per data volume and retention time)	Only information consumer is charged (e.g. per download)	Both provider and consumer are charged	Further details (e.g. prices)
JOUVE						it's depend of kind of the contract
CINES		X				
CRN/M2ISA						
IN2P3		X				
CRDO						
Paris/LACITO						
CNRTL/ATILF			x			
TELMA/ENC		X				Jusqu'ici, dans la plupart des cas, pas de contribution financière demandée ; pour les services qui seront rendus à des partenaires extérieurs au CRN (donc hors IRHT et EnC) nous allons désormais demander une participation aux frais. Pour les utilisateurs finals, le service est gratuit et le restera.

Customer support details. What are the channels for customers to report problems with the services, for instance unavailability of some scientific data (e.g. journal)?

<i>Company:</i>	<i>None</i>	<i>Support contact mail (list or private) and/or phone</i>	<i>Trouble ticket based workflow (either commercial or OpenSource support workflow, e.g. http://www.helpdeskpilot.com, http://www.somix.com, http://www.otrs.org or http://www.bmc.com)</i>	<i>Help-desk or call-center. Complementary to the workflow, a help-desk or call-center is usually put in place for first level support filtering of the user requests. First level support usually also maintains knowledge base and FAQ lists for the customers.</i>	<i>Other (please specify)</i>
JOUVE		x	X		
CINES		x	X	X	
CRN/M2ISA		x			
IN2P3		x	X		L'aide aux 3000 utilisateurs du centre s'effectue par un envoi de mail à user.support@cc.in2p3.fr . Un ticket est automatiquement crée suite à cet envoi. Le ticket est ensuite affecté aux experts concernés.
CRDO Paris/LACITO CNRTL/ATILF		x	X		
TELMA/ENC		x			Ce support sera amélioré dans les 2 années qui viennent.

<i>Company:</i>	Support coverage					<i>Other (please specify)</i>
	<i>Not defined</i>	<i>Office hours</i>	<i>Office hours + best efforts</i>	<i>24/7 standby service (on-call)</i>	<i>24/7 helpdesk coverage</i>	
JOUVE			x			24/7 for disponibility of the service
CINES			x			24/7 on-call service for hardware server support
CRN/M2ISA	X					
IN2P3					x	
CRDO Paris/LACITO CNRTL/ATILF			x			
TELMA/ENC			x			Hors des heures de bureau, la surveillance repose sur la bonne volonté des personnes responsables. Il est souhaitable qu'une surveillance continue soit organisée.

Computer center resources

Here follows some general questions about the computer centre where the Scientific Technical Information (STI) management services are hosted

How many other services (approximately), apart from the Human and Social Science services, are hosted by the computer center?				
<i>Company:</i>	<i>None. Only Human and Social Science service is hosted at the computer center</i>	<i>Also hosting some other minor services for the local employees (e.g. web servers or mail) but human and Social Science is the main service.</i>	<i><10 other large services are hosted in the computer center. Human and Social science is only one customer (not necessarily the largest)</i>	<i>>10 other large services are hosted in the computer center. Human and Social science is only one customer (not necessarily the largest)</i>
JOUVE				
CINES				X
CRN/M2ISA	x			
IN2P3				X
CRDO		X		
Paris/LACITO				
CNRTL/ATILF		X		
TELMA/ENC				X

How many computers (approximately) are hosted in the same data center as the Human and Social science services?					
<i>Company:</i>	<i>less than 10</i>	<i>10 - 100</i>	<i>100 - 1000</i>	<i>More than 1000</i>	<i>Other (please specify)</i>
JOUVE			x		the data center have more than 350 computers
CINES			x		
CRN/M2ISA		X			
IN2P3				x	
CRDO	x				
Paris/LACITO					
CNRTL/ATILF	x				
TELMA/ENC		X			7 serveurs et environ 80 machines sur 2 sites. La plateforme technique TELMA est en fait actuellement hébergée et gérée par la direction des NT de l'école, qui a en outre à sa charge la gestion du parc informatique et des applications de l'école (intranet ; Internet : portail institutionnel, nombreuses applications documentaires), ainsi que sa politique informatique.

Please give the approximate numbers of machines for each operating system (OS) type									
<i>Company:</i>	<i>windows</i>	<i>linux (please specify distribution)</i>	<i>solaris</i>	<i>AIX</i>	<i>HP-UX</i>	<i>Tru64 UNIX</i>	<i>IRIX</i>	<i>MacOS</i>	<i>Other (please specify)</i>
JOUVE	60	> 280 (RedHat, Fedora, Debian)	25						
CINES	50	60		11					
CRN/M2ISA									
IN2P3		1200 SL4 64 bits (equivalent RHEL 4)	150	40					
CRDO	0	1 CentOS	0	0	0	0	0	0	2 (NetBSD)
Paris/LACITO									
CNRTL/ATILF	3 (Windows Server 2003)	3 (Red Hat Enterprise Server)							
TELMA/ENC	70	10 (Debian sur serveurs, Ubuntu ailleurs)							

What is the approximate aggregate usable disk space hosted at the data center?					
Company:	0 to 10,000 GB (GigaBytes)	10 to 100,000 GB	100 to 1000,000 GB	More than 1000,000 GB	Other (please specify)
JOUVE			X		
CINES			X		
CRN/M2ISA			X		
IN2P3					x
CRDO	X				
Paris/LACITO					
CNRTL/ATILF	X				
TELMA/ENC			x		

Tape or other long term storage				
Company:	None	Magnetic Tape	Optical media, e.g. CD R/W, DVD	Other (please specify)
JOUVE		X		AIT3 used, in future LTO4
CINES		X		
CRN/M2ISA			X	disques externes amovibles
IN2P3		X		
CRDO Paris/LACITO	x			
CNRTL/ATILF		X		
TELMA/ENC		X		

What is the approximate aggregate usable tape (or other media) space hosted at the data center?					
Company:	None	Manually operated tape media space GB	Number of manual tape drives	Automated (robotic) tape media space GB	Number of automated drives
JOUVE				40000	10
CINES				150000	600 ⁴
CRN/M2ISA					
IN2P3				6000000	
CRDO	0	0	0	0	0
Paris/LACITO					
CNRTL/ATILF				6000	1
TELMA/ENC		1000	1		

What are the main robotic automation, tape media and drive types used at your data center?			
Company:	Main media type used, vendor and model (e.g. LTO-3, SDLT, Sun STK9940B, IBM 3592E, ...)	Tape automation solution, vendor and model (e.g. SUN STK SL8500)	Other
JOUVE		AIT3 used, in future LTO4	
CINES		STK9940	STK9310
CRN/M2ISA			
IN2P3		LTO-3, SUN STK 9940	hpss ⁵
CRDO		None	None
Paris/LACITO			
CNRTL/ATILF		LTO-2	Overland Serie 2000
TELMA/ENC		LTO-2	

⁴ The 600 drives reported by CINES was a misunderstanding and meant the number of tapes. The actual number of drives is 6

⁵ The answer 'hpss' is a misunderstanding of the question. HPSS is the Hierarchical Storage System (HSM) used at CC-IN2P3. The tape automation solution (tape robot) used is SUN STK SL8500 and STK9310

Media migration. Storage media deteriorate with time so in order to minimize risk of loss archived data should periodically be copied to new media. This is usually also attractive from a financial perspective since new media usually has higher density and better bandwidth.

Company: What is the media migration cycle at your data center? (years)

JOUVE	
CINES	5
CRN/M2ISA	
IN2P3	3
CRDO Paris/LACITO	None
CNRTL/ATILF	
TELMA/ENC	2

Power and cooling

<i>Company:</i>	<i>Available electrical power (kiloWatts)</i>	<i>Approximate current power consumption on full equipment load</i>	<i>Available cooling (kilowatts of heatload)</i>
JOUVE	400	200	
CINES	2500	300	1200
CRN/M2ISA			
IN2P3	2000	1000	400 ⁶
CRDO Paris/LACITO			
CNRTL/ATILF			
TELMA/ENC			

Staffing for computer center operation, system administration and service management.

<i>Company:</i>	<i>How many Full Time Equivalents (FTEs) are employed for the computer center operation?</i>	<i>How many Full Time Equivalents (FTEs) are employed for system administration tasks?</i>	<i>How many Full Time Equivalents (FTEs) are employed for managing the IT services (network, archive, applications, ...)</i>
JOUVE	21	10	10
CINES	10	10	20
CRN/M2ISA			
IN2P3	70	8	45
CRDO Paris/LACITO	1	1	1
CNRTL/ATILF	1	1	1
TELMA/ENC	6	1	4

What are the staff categories for operation, system administration and service management

<i>Company:</i>	<i>Computer center operation</i>			<i>System administration</i>		<i>Service management</i>		<i>Scientific staff (e.g. PhD students)</i>
	<i>IT technicians</i>	<i>IT engineers</i>	<i>Scientific staff (e.g. PhD students)</i>	<i>IT technicians</i>	<i>IT engineers</i>	<i>IT technicians</i>	<i>IT engineers</i>	
JOUVE	X	X		X	X	x	X	
CINES	X	X		X	X	x	X	X
CRN/M2ISA								
IN2P3	X	X		X	X	x	X	
CRDO Paris/LACITO		X			X		X	
CNRTL/ATILF	X				X		X	
TELMA/ENC	x	X	x		x	x	x	X

⁶ The current capacity of the CC-IN2P3 computer room is 1 MW for computing equipment, which corresponds to ~1.7 MW at full load including cooling. In other word, we have enough cooling power to cool 1 MW of computing equipment.

Staff coverage										
Company:	Computer center operation			System administration			Service management			Other (please specify)
	Office hours	24/7 shift	24/7 standby	Office hours	24/7 shift	24/7 standby	Office hours	24/7 shift	24/7 standby	
JOUVE	x	X		x			x			we have people 24/5 (not week end)
CINES	x	X		x			x			
CRN/M2ISA										
IN2P3	x		X	x		x	x			
CRDO	x			x			x			
Paris/LACITO										
CNRTL/ATILF	x			x			x			
TELMA/ENC	x			x			X			

IT budget in units of 1000 Euros			
Company:	Personnel	Material (including running maintenance)	Licenses
JOUVE			
CINES			
CRN/M2ISA			
IN2P3	2000		8000
CRDO Paris/LACITO	150		10
CNRTL/ATILF	100		10
TELMA/ENC			300

Please add here any other technical details (e.g. external bandwidth) of your data centre that you think could be relevant for this survey.

Company:	
JOUVE	We have two different connections (France Telecom and Neuf Cegetel) We upgrade regularly the links
CINES	3 x 2Gbits WAN Link (Renater) HPC 2,5 TFlops
CRN/M2ISA	Le système de sauvegarde du service informatique de la Fondation de la Maison des Sciences de l'Homme de Paris (FMSH) est en cours de création. Il sera opérationnel au mois de Mai 2008. Les serveurs et les ordinateurs ont des systèmes d'exploitation suivants : solaris sparc(sun), linux (mandriva, debian), microsoft, MacOS X.
IN2P3	
CRDO	None
Paris/LACITO	
CNRTL/ATILF	
TELMA/ENC	

Appendix B: Questionnaire pour les fournisseurs d'IST pour la communauté SHS en France

Veuillez donner le nom et une brève description de votre (ou vos) service(s) IST pour pour la communauté SHS	
Answer Options	Response Count
	5
<i>answered question</i>	5
<i>skipped question</i>	0

Number	Response Date	Response Text
1	11/26/2007 12:45:00	<p>Centre National pour la Numérisation de Sources Visuelles</p> <p>Le Centre National pour la numérisation de Sources Visuelle est un centre de ressources numériques (CRN) du CNRS créé en 2006 par le Département Sciences Humaines et Sociales et par la direction de l'information scientifique du CNRS.</p> <p>Il a pour mission la mise en œuvre d'une infrastructure d'informatisation des données dans le respect des standards et en favorisant les formats et outils libres, de méthodes de préservation des documents numériques, d'une aide à la diffusion de documents numériques visuels (photos, diapos, carnets de terrains, cartes, planches, dessins, croquis, etc.) pour des activités de recherche scientifique au travers de plateformes web 1.0 ou 2.0, d'extranets, d'entrepôt OAI-PMH, etc</p> <p>Il est adossé pour son fonctionnement au pôle histoire des sciences et des techniques en ligne du Centre Alexandre Koyré/CRHST et collabore avec le Très Grand Équipement ADONIS.</p>
2	11/28/2007 13:24:00	<p>Programme PERSEE</p> <p>Numérisation, documentation et diffusion de collections de revues SHS, du premier numéro jusqu'aux numéros récents (barrière mobile). Ce programme se caractérise par :</p> <ul style="list-style-type: none"> - respect du droit d'auteur - diffusion gratuite - collaboration avec les éditeurs - Open source et normes

3	12/11/2007 14:36:00	<p>Né en 1999, Revues.org est le plus ancien portail de revues en sciences humaines et sociales en France. Il est ouvert aux périodiques de qualité désireux de publier en ligne du texte intégral. Il est porté par le Centre pour l'édition électronique ouverte (CLEO), unité qui associe le CNRS, l'EHESS, l'Université de Provence et l'Université d'Avignon. Cette unité inscrit son action dans le cadre du Très grand équipement ADONIS.</p> <p>Revues.org accompagne déjà plus de 110 revues publiées par des sociétés savantes, des grands établissements de recherche, des presses universitaires et des éditeurs privés. La plupart ont également une édition papier. Elles relèvent des sciences humaines et sociales au sens large : histoire, sociologie, anthropologie, géographie, archéologie, linguistique, littérature, histoire de l'Art, économie, sciences politiques, philosophie... La personnalité de chaque revue est mise en valeur par une charte graphique spécifique. La revue conserve le contrôle de son édition électronique et son indépendance éditoriale.</p> <p>Le portail est enrichi par un ensemble d'outils documentaires qui font référence : Calenda, le calendrier des sciences sociales, L'Album des sciences sociales, répertoire raisonné de liens, et In-extenso, moteur de recherche des sciences humaines et sociales.</p>
4	12/14/2007 18:41:00	<p>CRN M2ISA : Centre de Ressources Numériques de données spatiales : "Méthodologies pour la Modélisation de l'Information Spatiale Appliquées aux sciences de l'homme et de la société"</p> <p>On entend par données spatiales ou géographiques des données qui sont géoréférencées, géolocalisées en longitude et en latitude ou dans un système de projection.</p> <p>Le CRN M2ISA est composé d'un Portail M2ISA hébergé par le service informatique de la FMSH et d'un GéoPortail M2ISA hébergé par le CRN M2ISA. Le Portail M2ISA a pour objectif le dépôt et le téléchargement</p> <ul style="list-style-type: none"> - des données spatiales et/ou thématiques accompagnées de leurs métadonnées conforme à la norme géographique internationale ISO 19115/TC 211 - d'outils comme des scripts, des programmes informatiques - de documentation - l'accès au géoportail M2ISA à partir du portail M2ISA - l'accès à des services cartographiques à partir du géoportail M2ISA. Un service cartographique est constitué d'information spatiale et/ou thématique dynamique géoréférencées, accessibles directement d'une façon dynamique. Un service cartographique N'EST PAS une collection de cartes numérisées.
5	01/08/2008 14:31:00	<p>Très liée au monde scientifique, EDP Sciences (Édition Diffusion Presse Sciences), filiale de la SFP (Société Française de Physique), de la Société Française de Chimie et de la Société de Mathématiques Appliquées et Industrielles et aussi partenaire d'autres sociétés savantes et institutions, participe à la communication et à la diffusion de la science vers les publics spécialisés (chercheurs, ingénieurs, étudiants, etc.) et non spécialisés (grand public, décideurs, éducation). EDP Sciences produit et publie des revues internationales, ainsi que des livres ou des sites Internet à dominante scientifique ou technique.</p>

En quelle année est-il mis en ligne?	
Answer Options	Response Count
	5
<i>answered question</i>	5
<i>skipped question</i>	0

Number	Response Date	Response Text
1	11/26/2007 12:45:00	2005
2	11/28/2007 13:24:00	janvier 2005
3	12/11/2007 14:36:00	Le portail Revues.org est né en 1999.
4	12/14/2007 18:41:00	Le Portail M2ISA a été mis en ligne en Décembre 2006
5	01/08/2008 14:31:00	1996

Est-ce que vous êtes à la fois fournisseur du service de gestion IST et centre de calculs qui héberge ce service (machines, stockage et archivage des données)?		
Answer Options	Response Percent	Response Count
Oui	0.0%	0
Non (indiquez svp les centres de calculs que vous utilisez)	100.0%	5
	<i>answered question</i>	5
	<i>skipped question</i>	0

Number	Response Date	Non (indiquez svp les centres de calculs que vous utilisez)
1	11/26/2007 12:45:00	Centre de Calcul de l'IN2P3-CNRS
2	11/28/2007 13:24:00	CINES
3	12/11/2007 14:36:00	Le CCSD nous fournit un espace dans leur baie de serveurs qui se trouve à l'IN2P3. Le STOCKAGE est assuré par le CRN M2ISA
4	12/14/2007 18:41:00	L'ARCHIVAGE est assuré par le service de ressources informatique (SRI) de la Fondation de la Maison des Sciences de l'Homme de Paris (FMSH)
5	01/08/2008 14:31:00	Société : Internet Fr Immeuble Odyssee 2, 12 chemin des Femmes 91300 MASSY 01 64 53 12 12

Quel type de données sociales et humaines de la Science (SHS) est géré par votre service IST?		
Answer Options	Response Percent	Response Count
Textes (articles, journaux, livres, ...)	100.0%	5
Images (photos, schémas, cartes, ...)	60.0%	3
Audio	0.0%	0
Vidéo	0.0%	0
Base de données locales de bibliographie	0.0%	0
Documents mélangés (manuscrits avec les images + le texte, dossier de collection de XML, ...)	60.0%	3
Source primaire numérisée (textuelle, iconographique, archéologique, etc.)	40.0%	2
Autre (indiquez svp)	20.0%	1
	<i>answered question</i>	5
	<i>skipped question</i>	0

The answer to 'Autre' was too long to fit in the table:

'- Charte d'utilisation détaillée des données spatiales - Textes juridiques concernant les données spatiales - Directives Européennes et internationales concernant les données spatiales - Base de Données Géographiques structurées et modélisées: DCW (Digital Chart World de 1993), SRTM (Modèles numériques de terrain obtenu par un radar stéréoscopique de l'USGS au niveau mondial avec une résolution de 90m), Réseaux de transport au niveau européen, - Systèmes d'Information Géographique sous forme de Géodatabase topologique - source primaire numérisée non géoréférencée : cartes récentes ou anciennes, photographies aériennes - source primaire numérisée géoréférencée : cartes récentes ou anciennes, photographies aériennes, - source primaire numérique géoréférencée : orthophotoplans, photographies aériennes, radar, satellites - source dérivée numérique : modèle numérique de terrain, modélisation hydrologique, T.I.N. - Remarques : Les cartes, les photographies aériennes, les orthophotoplans NE SONT PAS considérés comme des images. Ce sont des sources d'informations géographiques et thématiques à partir desquelles sont extraites les informations géographiques, puis les données spatiales et/ou thématiques'

Qui maintient le service IST et le logiciel associé (scripts php, interface aux bases de données, ...)?		
Answer Options	Response Percent	Response Count
Votre organisation	100.0%	5
Le centre de calculs en tant qu'élément des services qu'il vous fournit	0.0%	0
Autre (indiquez svp)	0.0%	0
	<i>answered question</i>	5
	<i>skipped question</i>	0

Volume total de données stockées (approximatif en gigaoctets)			
Answer Options		Response Percent	Response Count
Total aujourd'hui		100.0%	5
Augmentation annuelle		100.0%	5
		<i>answered question</i>	5
		<i>skipped question</i>	0
Number	Response Date	Total aujourd'hui	Augmentation annuelle
1	11/26/2007 12:45:00	2000	750
2	11/28/2007 13:24:00	20 000 Go (20 To)	7 000 Go
3	12/11/2007 14:36:00	40	5
4	12/14/2007 18:41:00	5000 gigaoctets (5 téras)	10000 gigaoctets (10 téras) en moyenne
5	01/08/2008 14:31:00	Environ 108 000 articles soit 170 Go de fichiers	2007/2006 : + 235% (production d'archives), 2005 : 20%

Comment le centre de calculs facture-t-il les ressources allouées à vos services de STI (machines, espace disc, archivage, personnel) ? Par exemple, payez-vous un montant annuel selon l'utilisation ou une somme fixe convenue d'avance pour une capacité et durée déterminée	
Answer Options	Response Count
	4
<i>answered question</i>	4
<i>skipped question</i>	1

Number	Response Date	Response Text
1	11/26/2007 12:45:00	Dans le cadre de l'équipement Mi-lourd du Laboratoire et en partenariat avec le Département SHS et le TGE ADONIS du CNRS, nous avons une convention d'hébergement avec le CC-IN2P3 pour 4 ans. PERSEE est piloté et financé par le Ministère (DGES/SDBIS).
2	11/28/2007 13:24:00	Une convention annuelle fixe le montant de l'hébergement et de la gestion du programme par le CINES
3	12/14/2007 18:41:00	- Pour l'instant gratuit. - L'archivage ne sera opérationnel qu'en Mai 2008. Suite à cela, la politique tarifaire changera.
4	01/08/2008 14:31:00	Forfait annuel

Quel type d'accès aux serveurs est exigé pour la gestion du service d'IST? Par exemple : la mise à jour des logiciels du service IST lui-même ; les changements de configuration et/ou bien d'autres modifications aux services de production.		
Answer Options	Response Percent	Response Count
Vous n'avez pas accès direct aux machines du centre de calculs. Chaque fois qu'un changement est exigé, vous devez soumettre la demande de changement qui va être effectué par le personnel du centre de calculs.	20.0%	1
Les machines de production sont sous la commande du centre de calculs. Vous avez l'accès administrateur aux machines de développement et de certification	20.0%	1
Vous avez l'accès limité (pas d'administrateur) aux machines, ce qui est suffisant pour une certaine gestion de service (par exemple : la vérification de logs)	20.0%	1
Le centre de calculs vous a accordé l'accès d'administrateur à toutes les machines	20.0%	1
Autre (indiquez svp)	100.0%	5
	<i>answered question</i>	5
	<i>skipped question</i>	0

Number	Response Date	Autre (indiquez svp)
1	11/26/2007 12:45:00	Il serait intéressant de développer la seconde utilisation au CC-IN2P3.
2	11/28/2007 13:24:00	Quelques scripts développés par le centre de calculs nous permettent également d'effectuer quelques commandes administrateur (relance du serveur apache par exemple)
3	12/11/2007 14:36:00	Nous avons besoin d'un accès root sur les machines et de maîtriser nous même les serveurs.
4	12/14/2007 18:41:00	Le centre de calcul a accordé l'accès d'administrateur pour la machine qui héberge le portail M2ISA
5	01/08/2008 14:31:00	Accès limité mais nous avons une machine supplémentaire identique à un serveur web et que nous administrons. Cette machine nous permet de tester les évolutions du site avant de les mettre en ligne sur les différents sites web afin de limiter les risques.

Veuillez indiquer qui est responsable des tâches d'administration du système suivant				
Answer Options	Votre équipe	Le personnel du centre de calcul	Autre	Response Count
Installation de machines	1	3	1	5
Application de mises à jour du noyau et OS (system d'exploitation)	1	3	1	5
La sécurité de la machine	1	4	0	5
La sécurité d'application d'IST	3	1	1	5
Monitoring de matériel et exécution de tâches courantes (par exemple nettoyage de /tmp)	3	2	0	5
Monitoring du service d'IST (vérifiant que le service fonctionne)	3	2	0	5
Archivage des données d'IST	1	3	1	5
Au cas où vous avez répondu 'Autre', veuillez svp détailler			2	
<i>answered question</i>				5
<i>skipped question</i>				0

Number	Response Date	Au cas où vous avez répondu 'Autre', veuillez svp détailler
1	11/28/2007 13:24:00	Autre = collaboration des deux équipes
2	12/14/2007 18:41:00	L'installation de machines et l'application de mises sont faites aussi bien par notre équipe que par le personnel du centre de calcul. Le système d'archivage ne sera opérationnel qu'en Mai 2008

Sur quel système de base de données repose votre service IST?		
Answer Options	Response Percent	Response Count
Aucun (pas de système de base de données)	0.0%	0
PostgreSQL	60.0%	3
MySQL	80.0%	4
GDBM (The GNU database manager)	0.0%	0
SAP maxDB	0.0%	0
Microsoft Access	0.0%	0
Microsoft SQL Server	0.0%	0
Oracle	0.0%	0
IBM DB2	0.0%	0
Autre (précisez)	40.0%	2
<i>answered question</i>		5
<i>skipped question</i>		0

Number	Response Date	Autre (précisez)
1	11/26/2007 12:45:00	XML file system
2	12/14/2007 18:41:00	ArcIMS

Qui effectue les tâches d'administration de base de données, comme (Répondez svp N/A si personne ne s'occupe de cette tâche)					
Answer Options	Votre équipe	Le personnel du centre de calculs	Autre	N/A (non applicable)	Response Count
Initialisation et configuration de base du système de bases de données	2	2	1	0	5
sauvegarde/récupération des données	1	4	0	0	5
validation de la stratégie et de l'implémentation de la sauvegarde/récupération des données	1	3	1	0	5
mise à jour, application des correctifs de sécurité	1	4	0	0	5
optimisation des paramètres du système de bases de données	1	3	1	0	5
optimisation des requêtes et des procédures stockées	3	1	1	0	5
placement des fichiers de bases de données sur le système de stockage afin de satisfaire aux besoins	2	3	0	0	5
Au cas vous avez répondu 'Autre', veuillez svp détaille					1
<i>answered question</i>					5
<i>skipped question</i>					0

Number	Response Date	Au cas vous avez répondu 'Autre', veuillez svp détaille
1	11/28/2007 13:24:00	Autre = collaboration des deux équipes

Utilisez-vous une solution haute disponibilité (HA) pour votre service de base de données?		
Answer Options	Response Percent	Response Count
Non	50.0%	2
Matériel spécial, par exemple un diskarray avec multiple connexion	0.0%	0
Solution HA propriétaire, par exemple Oracle DataGuard ou RAC	0.0%	0
Autre (précisez)	75.0%	3
<i>answered question</i>		4
<i>skipped question</i>		1

Number	Response Date	Autre (précisez)
1	11/28/2007 13:24:00	solution à l'étude, non déployée pour le moment
2	12/11/2007 14:36:00	HAProxy
3	01/08/2008 14:31:00	Nous utilisons plusieurs serveurs contenant les memes bases de données et un load balancer. Les serveurs de bases de données sont en HA linux (haute disponibilité) linux.

Est-ce que votre application de service IST...		
Answer Options	Response Percent	Response Count
nécessite des privilèges administrateurs dans la base de données (possibilité de créer des utilisateurs, des éléments de stockage...) durant l'installation	100.0%	5
nécessite des privilèges administrateurs dans la base de données (possibilité de créer des utilisateurs, des éléments de stockage...) durant le fonctionnement normal	20.0%	1
rétabli la connexion à la base de données de façon automatique en cas de dysfonctionnement du réseau ou de la base de données	40.0%	2
est fournie avec un script d'installation qui permet la création des objets dans la base de données (indexes, tables...)	80.0%	4
a ses mises à jour fournies avec des scripts de mise à jour d'une version à une autre (et aussi pour revenir en arrière)	60.0%	3
nécessite des opérations de maintenance régulière, veuillez préciser lesquelles ci-dessous	20.0%	1
	détails	1
		<i>answered question</i> 5
		<i>skipped question</i> 0

Number	Response Date	Details
1	12/11/2007 14:36:00	Suppression, création de bases Modifications de tables Optimisation du moteur de bases de données.

Dans les 5 ans à venir, quels sont les besoins qui vous semblent importants pour votre service IT		
Answer Options	Response Percent	Response Count
Augmenter le stockage	80.0%	4
Augmenter la performance du calcul	40.0%	2
Améliorer la disponibilité du service IST	100.0%	5
Autre (précisez)	40.0%	2
		<i>answered question</i> 5
		<i>skipped question</i> 0

Number	Response Date	Autre (précisez)
1	12/11/2007 14:36:00	Une salle blanche destiné aux SHS
2	12/14/2007 18:41:00	Augmenter le personnel et les crédits

Dans les 5 ans à venir, que voulez-vous développer ou changer au niveau d'application IST?		
Answer Options	Response Percent	Response Count
Développer plus d'applications de Web	80.0%	4
Changer vos applications de Web à d'autres technologies	20.0%	1
Autre (indiquez svp)	20.0%	1
		<i>answered question</i> 5
		<i>skipped question</i> 0

Number	Response Date	Autre (indiquez svp)
1	12/14/2007 18:41:00	Evolution vers ArcGIS Server

Êtes-vous satisfait des ressources, l'infrastructure et service fournis par le centre de calculs qui héberge votre service IST?		
Answer Options	Response Percent	Response Count
Très satisfait	60.0%	3
Satisfait	20.0%	1
Pas satisfait	20.0%	1
<i>answered question</i>		5
<i>skipped question</i>		0

Quels aspects étaient les plus importants dans la classification globale que vous avez établie dans la question précédente						
Answer Options	Non important	Important	Très important	N/A	Rating Average	Response Count
Les ressources (machines et stockage) allouées par rapport à la charge sur le service IST	0	2	3	0	2.6	5
Stabilité et disponibilité des machines et des ressources de stockage utilisées pour les services d'IST	0	0	5	0	3	5
La qualité de l'administration de système	1	0	4	0	2.6	5
Assistance en cas de problème pendant les heures de travail	1	0	4	0	2.6	5
Assistance en cas de problème en dehors des heures de travail	1	3	1	0	2	5
Flexibilité (ou manque de) en vous accordant accès aux machines pour des investigations sur les problèmes avec les services d'IST	0	3	2	0	2.4	5
Le coût courant pour héberger les services (et données) d'IST	0	2	0	3	2	5
<i>answered question</i>						5
<i>skipped question</i>						0

Quel ont été les problèmes principaux que vous avez rencontrés avec les services et ressources fournis par le centre de calculs ?						
Answer Options	Pas un problème	Parfois un problème	Souvent un problème	N/A	Rating Average	Response Count
Panne (par exemple. coupure de courant, problème de refroidissement, panne de réseau)	2	1	0	2	1.333333	5
nterruptions fréquentes mais prévues pour maintenance de ressources	2	1	0	2	1.333333	5
Interruptions fréquentes mais imprévues dues aux problèmes des ressources ou infrastructure	1	1	0	2	1.5	4
Archives incertaines (par exemple : perte ou corruption de données)	2	1	0	2	1.333333	5
Transfert instable ou lent d'information (téléchargement)	2	1	0	2	1.333333	5
Assistance en cas des problèmes	3	0	0	2	1	5
Autre (indiquez svp)						1
<i>answered question</i>						5
<i>skipped question</i>						0

Number	Response Date	Autre (indiquez svp)
1	11/28/2007 13:24:00	Sur 3 années, la moyenne des interruptions prévues ou non ne dépasse pas 5 interruptions par an

Comment avez-vous choisi le centre de calculs que vous utilisez?		
Answer Options	Response Percent	Response Count
Appel aux offres	20.0%	1
Collaboration existante	60.0%	3
Imposé par votre fournisseur de fonds	20.0%	1
Autre (précisez)		0
<i>answered question</i>		5
<i>skipped question</i>		0

Si, pour quelque raison, vous étiez obligés de changer le centre de calculs, quelles seraient les difficultés principales ?							
Answer Options	Problème principal	Problème potentiel	Inconvénience mais pas un problème	Pas un problème	N/A	Rating Average	Response Count
Déplacer les données	1	3	0	1	0	2.2	5
Déplacer le logiciel d'application de service de STI	1	2	2	0	0	2.2	5
Intégrer le service de STI dans la nouvelle infrastructure	0	0	2	1	1	3.333333	4
Maintenir l'accès pendant que le service est déplacé	1	2	2	0	0	2.2	5
L'absence du service pendant le mouvement	2	2	1	0	0	1.8	5
Formation du nouveau personnel pour courir vos services de STI	2	1	1	1	0	2.2	5
Autre (indiquez svp)							2
<i>answered question</i>							5
<i>skipped question</i>							0

Number	Response Date	Autre (indiquez svp)
1	12/11/2007 14:36:00	Migration des DNS.
2	12/14/2007 18:41:00	L'application développée est composée de 2 applications imbriquées l'une dans l'autre : celle du Portail M2ISA qui peut poser un problème potentiel et celle du GéoPortail qui, elle, pose un vrai problème principal.

Appendix C: End user Survey

Veuillez indiquer brièvement votre domaine de recherche (exemple: archéologie, géographie, ethnologie, histoire)	

Avez-vous déjà utilisé un portail Web pour accéder aux informations dans votre domaine scientifique (exemple: articles, journaux, images,...)?	
Answer Options	
Oui	
Non	

Quel type d'information accédez-vous habituellement?	
Answer Options	
Documents des textes (articles, journaux, livres,...)	
Images (photos, schémas, cartes,...)	
Audio	
Vidéo	
Bases de données locales de bibliographie	

Documents mélangés (manuscrits avec les images + le texte, dossiers de collection de XML,...)	
Source numérisées (textuelles, iconographiques, archéologiques, etc)	
Autre (indiquez svp)	

Veuillez indiquer les sites que vous avez utilisés					
Answer Options	Ne connais pas	Connais mais non approprié pour votre recherche	Approprié mais jamais utilisé	De temps en temps	Régulièrement
Persée (http://www.persee.fr/)					
Revue.org (http://www.revues.org/)					
TELMA (http://www.cn-telma.fr/)					
M2ISA (http://www.m2isa.fr/)					
CNRTL (http://www.cnrtl.fr/)					
CRDO (http://crdo.fr/ , http://crdo.up.univ-aix.fr/ , http://crdo.vjf.cnrs.fr/)					
CN2SV (http://www.cn2sv.fr/)					
IRHT (http://www.irht.cnrs.fr/)					
ARCHEOVISION (http://archeovision.cnrs.fr/)					
BiblioSHS (http://biblioshs.inist.fr/)					
BnF (http://www.bnf.fr/)					
CAIRN (http://213.161.196.111/accueil.php)					
Calenda (http://calenda.revues.org/)					
Catalogue Collectif de France (http://ccfr.bnf.fr/portailccfr/servlet/LoginServlet)					
Elsevier (http://france.elsevier.com/html/index.cfm?act=accueil)					
Erudit (http://www.erudit.org/)					
Gallica (http://gallica.bnf.fr/)					
Irevues (http://irevues.inist.fr/)					
l'Album des sciences sociales (http://album.revues.org/)					
SHMESP (http://shmesp.ish-lyon.cnrs.fr/)					
Menestrel (http://menestrel.in2p3.fr/)					
SUDOC (http://www.sudoc.abes.fr/)					
Autre 1 (indiquez svp les noms ci-dessous)					
Autre 2 (indiquez svp les noms ci-dessous)					
Autre 3 (indiquez svp les noms ci-dessous)					
Autre 4 (indiquez svp les noms ci-dessous)					
Autre (indiquez svp les noms)					

Pour le portail le plus utilisé ci-dessus, combien de fois (environ) accédez-vous ?	
Answer Options	
Moins d'une fois par semaine	
Au moins une fois par semaine mais moins d'une fois par jour	
Au moins une fois par jour	
Au moins 10 fois par jour	
Plus de 10 fois par jour	
Autre (indiquez svp)	

Indiquez svp pour les différents portails que vous avez utilisés comment vous évalueriez globalement leurs services. Utilisez le bouton de « N/A » (non applicable) pour les services que vous n'avez pas utilisés.					
Answer Options	Très mécontent	Mécontent	Satisfait	Très satisfait	N/A
Persée (http://www.persee.fr/)					
Reves.org (http://www.revues.org/)					
TELMA (http://www.cn-telma.fr/)					
M2ISA (http://www.m2isa.fr/)					
CNRTL (http://www.cnrtl.fr/)					
CRDO (http://crdo.fr/ , http://crdo.up.univ-aix.fr , http://crdo.vjf.cnrs.fr)					
CN2SV (http://www.cn2sv.fr/)					
IRHT (http://www.irht.cnrs.fr/)					
ARCHEOVISION (http://archeovision.cnrs.fr/)					
BiblioSHS (http://biblioshs.inist.fr/)					
BnF (http://www.bnf.fr/)					
CAIRN (http://213.161.196.111/accueil.php)					
Calenda (http://calenda.revues.org/)					
Catalogue Collectif de France (http://ccfr.bnf.fr/portailccfr/servlet/LoginServlet)					
Elsevier (http://france.elsevier.com/html/index.cfm?act=accueil)					
Erudit (http://www.erudit.org/)					
Gallica (http://gallica.bnf.fr/)					
Irevues (http://irevues.inist.fr)					
l'Album des sciences sociales (http://album.revues.org/)					
SHMESP (http://shmesp.ish-lyon.cnrs.fr/)					
Menestrel (http://menestrel.in2p3.fr/)					
SUDOC (http://www.sudoc.abes.fr/)					
Autre 1 (indiquez svp les noms ci-dessous)					
Autre 2 (indiquez svp les noms ci-dessous)					
Autre 3 (indiquez svp les noms ci-dessous)					
Autre 4 (indiquez svp les noms ci-dessous)					

Quels aspects étaient pour vous les plus importants pour votre classification globale ci-dessus?			
Answer Options	Non important	Important	Très important
Pertinence scientifique du site vis-à-vis votre recherche			
Qualité et la complétude de l'information scientifique fournie par le site			
Temps de réponse pour des recherches d'information			
Temps de réponse pour l'accès à l'information (par exemple téléchargement)			
Facilité de naviguer et utiliser le portail			
Assistance pendant les heures de travail			
Assistance en dehors des heures de travail			
Coût			
Autre (indiquez svp ci-dessous)			
Autre (indiquez svp)			

Quels ont été les problèmes principaux rencontrés en utilisant le service?			
Answer Options	Pas un problème	Parfois un problème	Souvent un problème
Panne du service (ou une certaine partie importante)			
Difficile à interroger (formuler une recherche) pour trouver une information particulière			
Long temps de réponse			
Indisponibilisé d'information, c.-à-d. l'information a été trouvée mais l'accès (par exemple, le téléchargement) échoue			
Information manquante, c.-à-d. le dépôt ne contient pas une information particulière que vous recherchez			
Livraison (par exemple, le téléchargement) instable ou lente			
Obtenir l'aide (support)			
Autre (indiquez svp)			

Comment avez-vous découvert l'existence des portails d'informations que vous utilisez?	
Answer Options	
par un moteur de recherche généraliste (par ex. Google)	
par navigation de site en site	
par un portail (svp indiquez lequel ci-dessous)	
par le bouche à oreille, de collègue à collègue	
par revue ou newsletter	
Autre (indiquez svp)	

Veillez donner vos suggestions afin d'améliorer les services en ligne concernant l'Information Scientifique et Technique (IST) pour votre communauté scientifique en France.

--

Si vous n'utilisez pas les portails en ligne de Web pour accéder à l'information scientifique pour votre recherche, quelle en est la raison?

Answer Options	
Vous n'en avez pas besoin	
Vous ne voyez pas comment il pourrait être utilisé pour l'information qui vous intéresse	
Vous ne connaissez pas de portail pour votre secteur scientifique	
Un portail existe mais vous ne savez pas l'utiliser	
Un portail existe mais les prix sont trop élevés	
Autre (indiquez svp)	

Si à l'avenir vous deviez utiliser un portail en ligne de Web pour accéder à l'information scientifique pour votre recherche, évaluez svp l'importance des conditions énumérées ci-dessous

Answer Options	Pas Important	Important	Très important
La disponibilité pendant des heures de travail			
La disponibilité en dehors des heures de travail			
Temps de réponse pour des recherches d'informations			
Temps de réponse pour l'accès à l'information (par exemple téléchargement)			
Facilité de naviguer et utiliser le portail			
Support pendant les heures de travail			
Support en dehors des heures de travail			
Coût			
Autre (indiquez svp ci-dessous)			
Autre (indiquez svp)			

References

- [1] IT Infrastructure Library, ITIL ®, <http://www.itil-officialsite.com/home/home.asp>
- [2] ISO/IEC 20000-1:2005 Information technology -- Service management -- Part 1: Specification, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=41332
- [3] ISO/IEC 20000-2:2005 Information technology -- Service management -- Part 2: Code of practice, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=41333
- [4] OAIS, CCSDS 650.0-B-1 Reference Model for an Open Archive Information System (OAIS), <http://public.ccsds.org/publications/archive/650x0b1.pdf> (English), [http://public.ccsds.org/publications/archive/650x0b1\(F\).pdf](http://public.ccsds.org/publications/archive/650x0b1(F).pdf) (French)
- [5] http://www.digitalpreservation.gov/formats/fdd/browse_list.shtml
- [6] OAI, Open Archives Initiative, <http://www.openarchives.org/>
- [7] OGF PE-RG, Open Grid Forum Preservation Environments Research Group, <https://forge.gridforum.org/sf/projects/pe-rg>
- [8] EGEE, Enabling Grids for E-science, <http://www.eu-egee.org/>
- [9] DILIGENT, A Digital Library Infrastructure on Grid Enabled Technology, <http://www.diligentproject.org/>
- [10] D4SCIENCE: DILIGENT for Science and the role of standards, http://portal.etsi.org/docbox/Workshop/200712_ECCONCERTATION/DILIGENT%20for%20Science-1.pdf
- [11] DRIVER, Digital Repository Infrastructure Vision for European Research, <http://www.driver-support.eu/en/index.html>. The details of the French involvement can be found at <http://www.driver-support.eu/en/national/france.html>
- [12] HAL, Hyper Article en Ligne, <http://hal.archives-ouvertes.fr/>
- [13] CLARIN, Common Language Resources and Technology Infrastructure, <http://www.clarin.eu/>
- [14] OGF HASS-CG, Open Grid Forum Humanities, Arts, and Social Sciences Community Group, http://www.ogf.org/gf/group_info/view.php?group=hass-cg
- [15] ARCHEOGRID, <http://archeovision.cnrs.fr/fr/archeogrid/index.htm>
- [16] Institut Des Grilles, <http://idgrilles.lal.in2p3.fr/>
- [17] Guy Wormser private communication