Key Concepts and Terms

Many of these concepts and terms are not yet well-defined and have considerable overlap. Our goal in providing these definitions is partly to provide some understanding of what the concepts and terms mean, but also to spark discussion among the Cohort about how, as a community, we should use these terms and how they relate to a library framework for e-research support. We encourage the Cohort to discuss and improve these definitions over the course of the Institute.

Research Life Cycle
The life cycle approach observes each stages of a process, to understand the overall process better. The research life cycle begins with the conception of a research project (hypothesis), continues through its methodology design and data collection, analysis, and finally publication and archiving of research outputs (e.g. articles, datasets, software, models, etc.). Understanding the research life cycle helps libraries identify who in involved and what information is produced or transformed during each phase of the project. For a more detailed explanation of this key concept, see e-Science and the Life Cycle of Research by Charles Humphrey.

e-Research
The term e-Research here refers to the use of information technology to support existing and new forms of scholarly research in all academic disciplines, including the humanities and social sciences. E-research encompasses computational and e-science, cyberinfrastructure and data curation. E-Research projects often make use of grid computing or other advanced technologies, and are usually data intensive, but the concept also includes research performed digitally at any scale. E-research is useful here as a way to bridge the concept of e-science to other fields such as social science and the humanities. Just as e-science applies large-scale computing to processing vast amounts of scientific research data, e-research could include studies of large linguistic corpuses in the humanities, or integrated social policy analyses in the social sciences.

Cyberinfrastructure
Cyberinfrastructure (or CI) describes research environments that support advanced data acquisition, data storage, data management, data integration,data mining, data visualization and other computing and information processing services distributed over the Internet beyond the scope of a single institution. In scientific usage, cyberinfrastructure is a technological strategy for efficiently connecting laboratories, data, computers, and people with the goal of enabling novel scientific theories and knowledge. The term “cyberinfrastructure” was coined in the U.S. and other countries have different terms for this type of technological infrastructure. Cyberinfrastructure now often includes systems for managing, archiving and preserving data, in addition to data processing, and so can include digital libraries and archives and the software and hardware to support them.

Computational Science
Computational science (or scientific computing) involves constructing mathematical models and using quantitative analysis techniques and computers to analyze and solve scientific problems. Typically, it involves the application of computer simulations or other forms of computation to scientific problems. Scientists and engineers develop computer software to model systems being studied and run these programs with various sets of input parameters. The models often require very large-scale computation (see Grid Computing and High Performance Computing) and are often executed on supercomputers or distributed computing platforms. Computational science is distinct from computer science (the study of computation, computers and information processing) and is different from the traditional theoretical and experimental scientific methods. Example problems include modeling of ecosystems or economies, biological pathways,

e-Science

E-Science is computationally intensive science carried out in highly distributed network environments, such as science that uses immense data sets requiring grid computing or High Performance Computing to process. The term sometimes includes technologies that enable distributed collaboration, such as the Access Grid, and is sometimes used as an alternative term for Cyberinfrastructure (e.g. e-Science is the preferred term in the UK). Examples of e-Science research include data mining, and statistical exploration of genome and other –omic structures.

High Performance Computing
High Performance Computing (HPC) involves parallel-processing computers and programs used for scientific research or computational science. High-Performance Technical Computing (HPTC) generally refers to the engineering applications of cluster-based computing such as computational fluid dynamics and the building and testing of virtual prototypes. In recent years HPC systems have shifted from supercomputing architectures to computing clusters and grids.

Grid Computing
Grid computing refers to a combination of computer resources from multiple organizations to form a shared, integrated computing platform. The grid is a distributed system with scheduled (non-interactive) workloads involving large number of files. What distinguishes grid computing from conventional High Performance Computing systems such as cluster computing is that grids tend to be more loosely coupled, heterogeneous, and geographically dispersed. Although a grid can be dedicated to a specialized application, it is more common that a single grid will be used for a variety of different purposes. Grids are often constructed with the aid of general-purpose grid software known as middleware to create a “super virtual computer” composed of many networked loosely coupled computers acting together to perform very large tasks. Furthermore, “distributed” or “grid” computing, in general, is a special type of parallel computing that relies on networked computers. This is in contrast to the traditional notion of a supercomputer, which has many locally-connected processors.

Semantic Web
The Semantic Web is a "web of data" intended to enable computers to understand the semantics, or meaning, of information on the World Wide Web. It extends the network of hyperlinked human-readable web pages that currently populate the Web by publishing machine-readable metadata about Web documents and how they relate to each other. The Semantic Web is being promoted by the World Wide Web Consortium ("W3C"), which oversees the development of proposed Semantic Web standards. The “Semantic Web" is implemented with a set of formats and technologies including the Resource Description Framework (RDF), a variety of data interchange formats (e.g. RDF/XML, N3, Turtle, N-Triples), and notations such as RDF Schema (RDFS) and the Web Ontology Language (OWL), all of which are intended to provide a formal description of concepts, terms, and relationships within a given knowledge domain. More recently, the term “Linked Data” or “Linked Open Data” has become the preferred term for this concept of a data Web.

Ontology
In computer science and information science, an ontology is a formal representation of knowledge as a set of concepts within a domain, and the relationships between those concepts. An ontology is meant to be a "formal, explicit specification of a shared conceptualisation". It defines a shared vocabulary and taxonomy to model a domain — the definition of objects and/or concepts and their properties and relations. Ontologies are structural frameworks for organizing information and are used in artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science and information architecture as a form of knowledge representation about the world or some part of it.

Data Life Cycle
The data lifecycle – and data lifecycle management – deals with tracking, managing, and understanding data and metadata as it flows through organizations.

Data Curation
Data curation refers to the value-added activities and features that stewards of digital content engage in to make digital content meaningful or useful. The data portion of this term sometimes refers specifically to research data (the outcomes of conducting research) and sometimes to digital content of any kind.

Digital Curation
Digital curation includes *digital* preservation data *curation*, as initially defined by the Digital Curation Center of the UK when it was founded. This term encompasses the full lifecycle of digital content management: selection, preservation, maintenance, collection and archiving of digital assets. Digital curation is generally referred to the process of establishing and developing long term repositories of digital assets for current and future reference by researchers, scientists, historians, and scholars.

Data Provenance
A term most often used to describe the semantic meaning of a dataset, or what the library community usually calls metadata, as well as where the data came from and by what methods (closer to the archives concept of provenance).

Digital Preservation
Digital preservation is the set of processes and activities that ensure continued access to information and all kinds of records, scientific and cultural heritage existing in digital formats, and is an ongoing process. This includes the preservation of materials resulting from digital reformatting (e.g. scanning from print), but particularly information that is born-digital and has no analog counterpart. Digital preservation is defined as: long-term, error-free storage of digital information, with means for retrieval and interpretation, for the entire time span the information is wanted. Long-term is defined as "long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community". Interpretation means that the retrieved digital files can be decoded and transformed into usable representations that are meaningful to the user (human or computer). There is an important distinction between "bit preservation", i.e., preserving the original files intact and unchanged over time) and "functional preservation", or preserving the information in the files by means of reformatting, added documentation, or other processes that will enable users to interpret the information in the future.

Data Management Plans
Data Management Plans describe how a research project will manage, dissemination and share research results. This includes descriptions of the types of data, samples, physical collections, software, curriculum materials, and other materials produced by a project; the standards used for data and metadata; policies for access and sharing data, including privacy protection of privacy and intellectual property, and security; policies and provisions for re-use, re-distribution, and the production of derivatives; and methods of archiving data, samples, and other research products, and for preservation of access to them. Funding bodies increasingly require grant-holders to produce and maintain Data Management Plans, both at the proposal stage and during the project.

Data Governance
Data governance is the system of decision rights and accountabilities for information-related processes that describe who can take what actions with what information, and when, under what circumstances, using what methods. It includes strategies for data quality control, data management, data policies, business process management, and risk management for data in the context of an organization. It is the set of processes to insure that important data assets are formally managed throughout an organization (including virtual organizations such as large, international research collaborations). Data governance ensures that data can be trusted and that people can be made accountable for any adverse event affecting the data.

Data Visualization
Data visualizations are visual representations of data, or abstract information. In the context of e-Science data visualization is closely related to scientific visualization, an interdisciplinary branch of science primarily concerned with the visualization of three dimensional phenomena (architectural, meteorological, medical, biological, etc.), where the emphasis is on realistic renderings of volumes, surfaces, illumination sources, and so forth, perhaps with a dynamic (time) component. Visualizations and simulations are a key part of scientific communication in the digital era, and require sophisticated software to execute (i.e. the visualizations are often not static files that can be captured and preserved like a digital image).

Team Science
Solving big problems or "grand challenges" in science generally requires big teams, big budgets, and a long time frame. It often involves the collaboration of many different scientists and engineers from a wide

variety of disciplines in the context of a research center or institute, which also often attempts to integrate research with education, technology transfer efforts, outreach activities, and diversity enhancement programs. Team science requires elaborate mechanisms for cross-disciplinary communication and tools for collaboration and sharing.

Strategic Agenda
A strategic agenda is a set of coherent and aligned strategies that an organization identifies and which serve as an organization's platform for a strategic plan. The strategic agenda a library develops will guide implementation over the course of a few years as it is translated into the organization's strategic plan. For further information, see the Association of Research Libraries, "Chapter 4: Applying the Scenarios to Create Strategy" in *The ARL 2030 Scenarios: A User's Guide for Research Libraries*, (2010) pp. 41-46.
NOTE: It is not required that you adopt a scenarios approach to developing your strategic agenda as discussed in the ARL guide, rather the document is provided as a reference.