Biochimica et Biophysica Acta 1844 (2014) 2–11

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

The amino acid's backup bone — Storage solutions for proteomics facilities ☆,☆☆

Hagen Meckel [a,1], Christian Stephan [a,b,1], Christian Bunse [a], Michael Krafzik [a], Christopher Reher [a], Michael Kohl [a], Helmut Erich Meyer [a], Martin Eisenacher [a,*]

[a] Medizinisches Proteom-Center, Ruhr-Universität Bochum, Universitätsstrasse 150, D-44801 Bochum, Germany
[b] Kairos GmbH, Universitätsstrasse 136, D-44799 Bochum, Germany

ARTICLE INFO

Article history:
Received 10 December 2012
Received in revised form 14 May 2013
Accepted 15 May 2013
Available online 27 May 2013

Keywords:
Bioinformatics infrastructure
Information technology
Data archive
Data backup
Storage attached network
Storage redundancy

ABSTRACT

Proteomics methods, especially high-throughput mass spectrometry analysis have been continually developed and improved over the years. The analysis of complex biological samples produces large volumes of raw data. Data storage and recovery management pose substantial challenges to biomedical or proteomic facilities regarding backup and archiving concepts as well as hardware requirements. In this article we describe differences between the terms backup and archive with regard to manual and automatic approaches. We also introduce different storage concepts and technologies from transportable media to professional solutions such as redundant array of independent disks (RAID) systems, network attached storages (NAS) and storage area network (SAN). Moreover, we present a software solution, which we developed for the purpose of long-term preservation of large mass spectrometry raw data files on an object storage device (OSD) archiving system. Finally, advantages, disadvantages, and experiences from routine operations of the presented concepts and technologies are evaluated and discussed. This article is part of a Special Issue entitled: Computational Proteomics in the Post-Identification Era. Guest Editors: Martin Eisenacher and Christian Stephan.

1. Introduction

In recent years proteomic techniques for high-throughput identification and quantification of peptides and proteins have been steadily improved and have become very popular and significant in biomedical and other fields of research. Particularly the core technology liquid chromatography (LC) followed by mass-spectrometry (MS) is widely used in proteomics and is established as one of the state-of-the-art technologies for analyzing complex biological samples [1]. In the past decade from MS device to MS device generation [2,3], raw file sizes significantly grew due to two main reasons: better peptide LC separation comprising increasing resolution power (i.e. peak capacity) leads [4] to a higher amount of MS/MS-spectra; and the LC coupled mass spectrometers with higher resolution (up to 120,000 and slightly above) as well as faster scanning speed (up to 50 species/s) amplified this effect [5]. High throughput MS output of complex biological samples produces large volumes of raw data, up to gigabyte file size per run. This results in tremendous challenges for handling the data flood regarding storage and reliability in proteomic or biomedical research facilities [6]. Similarly, genomic research areas, with the recently established high throughput technologies such as next-generation sequencing (NGS) methods, produce comparable or larger amounts of raw data. This issue is connected to enormous demands on data storage requirements which also have to be well contemplated and considered before utilizing this technology [7]. Research facilities are often overwhelmed with managing such large volumes of data because of unawareness of backup concepts. In addition, these facilities often do not know the intricacies of backup/archive concepts and their interaction with hardware technologies [8]. In our facility we experienced this data growth (per run) over the last years (see Fig. 1) and built up backup and archiving solutions with low- to high-end storage technology that allow management of huge amounts of data and constantly increasing files.

Abbreviations: CD, compact disk; CIFS, common internet file system; CPU, central processor unit; DVD, digital versatile disk; FC, fiber channel; GUI, graphical user interface; GZ, gzip; HEP, high-energy physics; HTML, hypertext markup language; iRODS, integrated Rule-Orientated Data System; iSCSI, internet small computer system interface; JOOQ, java object orientated query; LTO, linear tape open; MSDX, mass spectrometry DX client; NAS, network attached storage; NFS, network file system; NGS, next-generation sequencing; OS, operating system; RAID, redundant array of independent disks; RAM, random-access memory; SAN, storage area network; SCSP, simple content storage protocol; SDK, software development kit; SMB, server messages block; SWT, standard widget toolkit; TCP/IP, transmission control protocol/internet protocol; USB, universal serial bus; UUID, universally unique identifier; VMM, virtual machine manager; WORM, write once read many

☆☆ This article is part of a Special Issue entitled: Computational Proteomics in the Post-Identification Era. Guest Editors: Martin Eisenacher and Christian Stephan.

* Corresponding author at: Working group Bioinformatics/Biostatistics, Medizinisches Proteom-Center, building ZKF E.043, Ruhr-Universität Bochum, D-44801 Bochum, Germany. Tel.: +49 234 32 29288; fax: +49 234 32 145553.
E-mail address: martin.eisenacher@rub.de (M. Eisenacher).
[1] These authors contributed equally to this work.

1570-9639/$ – see front matter © 2013 The Authors. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbapap.2013.05.018

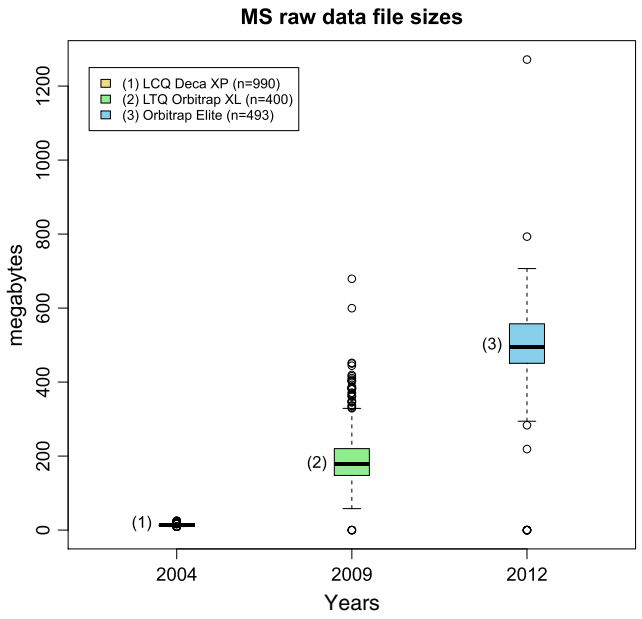

Fig. 1. Boxplots of complex samples (120 min runs) of MS raw files from three different mass spectrometry devices. The raw files were generated in the year 2004 (LCQ™ Deca XP), 2009 (LTQ Orbitrap XL) and 2012 (Orbitrap Elite). The average file size (n = number of files analyzed) increased significantly over the years, mainly due to better peptide LC separation and higher and faster scanning speed.

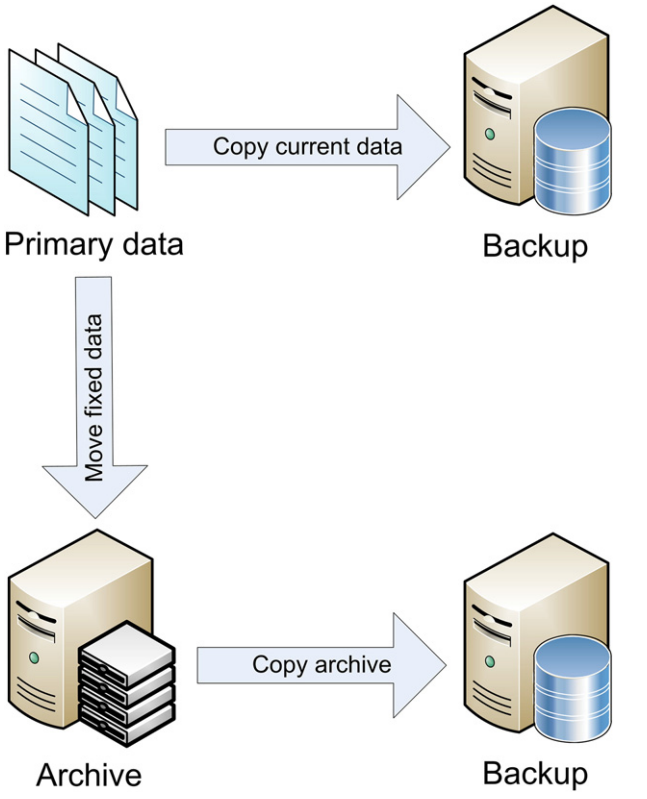

Fig. 2. Difference between backup and archive: A backup creates a secondary copy of primary data intended for recovery of current data and is generally overwritten periodically (weekly, monthly). Data can be restored from a specific date. The archiving process typically moves fixed data out of the active data workflow to secure unchangeable content in long-term preservation. Archived data should also be backed up or stored redundantly.

First of all, knowing the differences between backing up data and archiving it is essential when creating data management strategies to ensure long-term persistence, reliability and recoverability of high valuable raw data.

The terms "backup" and "archive"/"archiving" are widespread. Though both terms are widely used synonymously their meaning is different (see Fig. 2). Both terms represent different concepts for data management one should be aware of.

1.1. Backup

Let "primary data" be the active current data, which is in daily use and may change now and then. The main reason for performing a backup is to create copies of this primary data. The backup process makes point-in-time copies of files and folders usually to an alternative independent disk area. The backup represents a data copy of a specific date and time. It can be used for recovery when primary data is lost, destroyed or corrupted due to software or hardware problems or user failure. As primary data is subject to constant changes and modifications, previous backups become obsolete or out-of-date. For this reason backups should be repeated periodically, either manually or automatically by time scheduled software. Copying data from source to destination may follow one or more backup strategies which largely differ in disk space and execution time:

- Full Backup — contains all data
- Differential Backup — contains all new or changed data since full backup
- Incremental Backup — contains all new or changed data since last incremental backup.

1.2. Archive

Archiving data is used for long-term preservation and securing of unchanged, rare accessed or inactive data. This archivable primary data is generally not copied but moved from primary storage to an archival media destination (online or offline archive). Usually no primary data remains on the primary storage. This places great demands on archive technology to additionally assure immutability and reliability of the original data for the entire data life cycle. Archiving data helps to reduce the amount of data which normally needs to be put into the backup cycle. It is not unusual for archiving systems to provide additional metadata, which is attached to the archived content for reasons of documentation and faster search- and availability.

Backups should be performed for archived data as well, since the archive may – as may the primary data – become damaged. Since an archive usually contains unchanged data, one copy is sufficient. Alternatively the archive can be stored in an error-redundant system.

These concepts have been focused on already at a very early stage, and most organizations, for which data availability and reliability are crucial have established them. Among others, dentists and a waste company published their experiences with backup strategies and systems [9–12]. In many image processing facilities, technologies such as "automated tapes", "data grid" strategies and other backup and archiving concepts have been described and discussed [13–15]. In the field of genomics regarding next-generation sequencing (NGS), data management systems have been established, providing rule-based approaches for data replication and data protection [16]. In science fields of high-energy physics (HEP), data grid based or hierarchical storage strategies for handling large distributed data sets have been introduced and implemented [17,18]. Hospitals, medical and healthcare environments have to ensure data disaster recovery and availability and are bound by legal guidelines [19,20]. In the field of proteomics, data sharing in public repositories is discussed, while focusing rather on the results of proteomics data analysis (e.g. protein and peptide identification) than on "simple" archiving of raw data [21–26]. Generally speaking, laboratory environments are advised to follow recommendations of "good scientific practice". In this context for instance, the German Research Foundation proposed that "primary