Finally, we certainly agree that the PDB should have a standard data representation, although a well-designed ontology plays less of a role than that championed by King and his coauthors. Even so, the legacy requirements of our community and the dynamics of changing a global resource require that it be developed over time and in collaboration with our diverse user base. Anything less is a gross underestimation of the current usage and impact of PDB data.

COMPETING INTERESTS STATEMENT
The authors declare no competing financial interests.

Helen M Berman[1], Kim Henrick[2], Haruki Nakamura[3], John Markley[4], Philip E Bourne[5] & John Westbrook[6]

[1]RCSB Protein Data Bank, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [2]Macromolecular Structural Database, European Bioinformatics Institute, EMBL, Outstation, Hinxton, Cambridge, UK. [3]PDBj, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan. [4]BioMagResBank, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. [5]RCSB Protein Data Bank, San Diego Supercomputer Center and the Skaggs School of Pharmacy and Pharmaceutical Sciences at the University of California, San Diego, La Jolla, California 92093, USA. [6]RCSB Protein Data Bank, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. Correspondence should be addressed to H.M.B. e-mail: berman@rcsb.rutgers.edu

1. Berman, H.M., Henrick, K. & Nakamura, H. *Nat. Struct. Biol.* **10**, 980 (2003).
2. Berman, H.M., Henrick, K. Nakamura, H. & Markley, J.L. *Nucleic Acids Res.* **35**, D301–D333 (2007).
3. Fitzgerald, P.M.D. *et al.* in *International Tables for Crystallography* (eds. Hall, S.R. & McMahon, B.) 295–443 (Springer, Dordrecht, The Netherlands, 2005).
4. Westbrook, J., Henrick, K., Ulrich, E.L. & Berman, H.M. in *International Tables for Crystallography* (eds. Hall, S.R. & McMahon, B.) 195–198 (Springer, Dordrecht, The Netherlands, 2005).
5. Hall, S.R., Allen, A.H. & Brown, I.D. *Acta Crystallograph.* **A47**, 655–685 (1991).
6. Westbrook, J., Ito, N., Nakamura, H., Henrick, K. & Berman, H.M. *Bioinformatics* **21**, 988–992 (2005).
7. http://www.omg.org/cgi-bin/doc?lifesci/00-02-02
8. Westbrook, J. & Bourne, P.E. *Bioinformatics* **16**, 159–168 (2000).

Amanda C Schierz, Larisa N Soldatova & Ross D King respond:

The first point we would like to make is that we are very pleased to learn of the project to clean up the data in the PDB and to see on the website (http://www.wwpdb.org/) 'Announcement: Release of Remediated PDB Data' (16 April, 2007). This is a welcome development for the structural biology community.

Even so, we are disappointed with the reply from the wwPDB group. What we had hoped to read was a plan for structural biology to regain its lead in scientific data standards. Instead what the letter consists of is a series of red herrings, excuses for past problems, a complacent description of the current situation and a vague promise of jam tomorrow. Our main claims that mmCIF is a poor ontology and that the RCSB is a poor relational database are not seriously disputed.

Considering the red herrings: the wwPDB authors object to our use of "PDB" and of the term "Brookhaven Protein Data Bank" for post-1998 data. Yet, the title 'Overhauling the PDB' was *Nature Biotechnology*'s editorial suggestion, not ours, indicating that PDB is a generally accepted term for their organization. And although we should have deleted the word "Brookhaven" (which was erroneously introduced by editors at the proof stage), one must ask, 'What's in a name?' Would your data smell any sweeter with the correct name?

The wwPDB group also claims that we argue "that data organization in our data dictionary or any domain dictionary for that matter, should dictate the logical and physical organization of our database systems." We don't claim this. We are well aware of the differences between a logical and physical database model, which is why we were surprised that the RCSB PDB logical and physical model are exactly the same! The question of how an ontology can contribute to database design is an active area of research with high potential. The worry is that given the poor example of the RCSB PDB, database developers may draw the wrong conclusion about the usefulness of ontologies.

Considering the excuses: the wwPDB group argues that it has a lot of complex data to deal with. We, of course, accept this. But the problem is not new and PDB/wwPDB have had over 35 years to get things right. We have examined the wwPDB remediated chemical component dictionary and note that the obsolete component codes are now clearly labeled as such. Even so, we believe that some, perhaps many, of the mmCIF files on the remediated wwPDB FTP (file transfer protocol) site still contain incorrect data. For example, the nuclear magnetic resonance–obtained protein structures 1AXJ and 2FN2 are both supposedly remediated, yet both contain the mmCIF CELL category (which is defined as 'Data items in the CELL category record details about the crystallographic cell parameters').

The wwPDB also seem to put the blame for poor features in mmCIF on the International Union of Crystallography (IUCr). This seems to be an abdication of responsibility. Their claim that mmCIF was an ontology when Westbrook and Bourne[1] was written, but the meaning of the term 'ontology' has changed, is also weak.

To conclude, we hope that before the end of the decade, the wwPDB will present the structural biology community with guaranteed clean and self-consistent structural data, a state of the art ontology to represent these data and link it with other types of data and (at least) one state-of-the-art relational database to store and access the data.

1. Westbrook, J. & Bourne, P.E. *Bioinformatics* **16**, 159–168 (2000).

The Metabolomics Standards Initiative

To the editor:

The standards papers that *Nature Biotechnology* hosted online as part of a community consultation (http://www.nature.com/nbt/consult/index.html), in particular those by the Human Proteome Organization Proteomics Standardization Initiative (HUPO-PSI)[1,2] and the Functional Genomics Experiment (FuGE)[3] working groups, represent an important first step toward permitting the sharing of high-quality, structured data. We particularly applaud the open consultation solicited by *Nature Biotechnology* and advocate the early-community-involvement approach taken by HUPO-PSI, FuGE and the other working groups in the development of such guidelines and standards. These are the most effective ways to ensure that the output generated is pragmatic and the standards are both useful and widely accepted by the community.

As representatives of the nascent Metabolomics Standards Initiative (MSI)[4], we are following closely the work of the FuGE and the PSI working groups, leveraging on their work where commonality exists, such as the mass

Nature Biotechnology's online community consultation initiative (http://www.nature.com/nbt/consult/index.html) is intended to encourage researchers to participate in the development of guidelines/standards.

spectrometry and the sample preparation domains. MSI combines and thereby strengthens several preexisting groups and initiatives (including Standard Metabolic Reporting Structure (SMRS), ArMet and MIAMET)[5–7] in a concerted effort under the aegis of the Metabolomics Society[8]. As with other functional genomic approaches, we envisage a great deal of commonality in terms of experimental description.

The MSI working groups have drafted a series of manuscripts[9], outlining the work to date, and we intend to work closely with PSI working groups towards the development of common or interoperable standards. It is our view that reporting standards (checklists), syntax (format) and semantics (controlled vocabulary or ontology) should be reused across the functional genomics and systems biology standards communities, where applicable. This would benefit the entire scientific community by facilitating publication and dissemination of the results and simplifying the job of data integration[10]. From a technical perspective, it will be necessary to both remove redundancies and fill gaps between the domains that are covered by checklists, exchange formats and terminologies developed. These are certainly difficult but not insurmountable tasks and FuGE, developed and endorsed by many communities to 'unify' the exchange formats, is the first good exemplar project.

We would also like to highlight that the sociological barriers involved in these large-scale open standards efforts, such as PSI, can be extremely challenging, and thus will require extensive liaison between communities. With our experience in MSI, we are aware that managing this process of consensus building from start to finish takes time, resources and expertise. The time invested in these efforts to build commonalities and synergies among initiatives (e.g., between PSI and MSI) is often little, or at least not as continuous as it should be, due to lack of resources.

For these reasons, we express our appreciation for the efforts that each individual has put into the work behind FuGE and PSI guidelines. We take this opportunity also to encourage individuals or groups to join these initiatives, bringing with them their requirements, suggestions or critiques, and contributing to the development process in a constructive manner.

COMPETING INTERESTS STATEMENT
The authors declare no competing financial interests.

MSI Board Members: Susanna-Assunta Sansone[1], Teresa Fan[2], Royston Goodacre[3], Julian L Griffin[4], Nigel W Hardy[5], Rima Kaddurah-Daouk[6], Bruce S Kristal[7], John Lindon[8], Pedro Mendes[3,9,10], Norman Morrison[9,11], Basil Nikolau[12], Don Robertson[13], Lloyd W Sumner[14], Chris Taylor[1,11], Mariët van der Werf[15], Ben van Ommen[15,16] & Oliver Fiehn[17]

[1]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. [2]Department of Chemistry, University of Louisville, 2320 South Brook Street, Louisville, Kentucky 40292, USA. [3]School of Chemistry and Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester M1 7DN, UK. [4]Department of Biochemistry, Tennis Court Road, University of Cambridge, Cambridge, CB2 1QW, UK. [5]Department of Computer Science, University of Wales,

Penglais, Aberystwyth, Ceredigion, SY23 3DB, UK. [6]Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, North Carolina 27710, USA. [7]Department of Neurosurgery, Brigham and Women's Hospital, 221 Longwood Avenue, LM322B, Boston, Massachusetts 02115, USA. [8]Imperial College London, Department of Biomolecular Medicine, Sir Alexander Fleming Building, Exhibition Road, South Kensington, London SW7 2AZ, UK. [9]School of Computer Science, Kilburn Building, University of Manchester, Oxford Road, Manchester M13 9PL, UK. [10]Virginia Bioinformatics Institute, Virginia Tech, Washington Street MC 0477, Blacksburg, Virginia 24061, USA. [11]NERC Environmental Bioinformatics Centre, Oxford Centre for Ecology and Hydrology, Mansfield Road, Oxford, OX1 3SR, UK. [12]Iowa State University, 2210 Molecular Biology Building, Ames, Iowa 50011-1061, USA. [13]Molecular Profiling, Pfizer Global Research and Development, 2800 Plymouth Road, Ann Arbor, Michigan 48105, USA. [14]The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, Oklahoma 73410, USA. [15]TNO Quality of Life, PO Box 360, 3700 AJ Zeist, The Netherlands. [16]European Nutrigenomics Organization (NuGO), Bomenweg 2, 6703 HD Wageningen, The Netherlands. [17]University of California, Davis, Genome Center, 1 Shields Avenue, Davis, California 95616, USA.
e-mail: sansone@ebi.ac.uk

1. http://www.nature.com/nbt/consult/pdf/Taylor_et_al.pdf
2. http://www.nature.com/nbt/consult/pdf/Taylor_et_al_2.pdf
3. http://www.nature.com/nbt/consult/pdf/Jones_et_al.pdf
4. http://msi-workgroups.sourceforge.net
5. Lindon, J.C. *et al. Nat. Biotechnol.* **23**, 833–828 (2005).
6. Jenkins, H. *et al. Nat. Biotechnol.* **22**, 1601–1606 (2004).
7. Bino, R.J. *et al. Trends Plant Sci.* **9**, 418–425 (2004).
8. http://metabolomicssociety.org/
9. Flehn, O. *et al. Metabolomics* **3**, (3), doi: 10.1007/s11306-007-0070-6 (2007).
10. Field, D. and Sansone, S.A. *OMICS* **10**, 84–93 (2006).

Dichotomy in post-genomic microbiology

To the editor:

Your editorial in November (*Nat. Biotechnol.* **24**, 1299, 2006) discusses several initiatives and common 'platforms' that are being established to improve scientific communication and data comparison, including several standards under development, such as those for the analysis of microarray data[1]. We wish to raise a related concern about the unintentional development of a dichotomy in bacterial nomenclature in post-genomic microbiology, where strains for which the genome is known (sequenced strains) are increasingly treated as exemplary for the species. In addition, incomplete or incorrect bacterial names frequently occur in the genomic databases and literature. This has lead to the accumulation of unchecked information and the establishment of a parallel standard in microbiology, where sequenced strains are becoming the reference point instead of type strains.

The bacterial names associated with the 319 complete and published genome sequences, representing 232 different taxa, reported in GOLD database (http://www.genomesonline.org/; as of June 2007) were analyzed and compared with the designation reported in GenBank (http://www.ncbi.nlm.nih.gov/) for the same sequences (see **Supplementary Table 1** online). This evaluation revealed several inaccuracies (data reported refer to GOLD database).

First, in 11 cases only the genus name is given; to make matters worse, in only seven of these cases is the genus name valid. Second, for the remaining 308 strains, 18 names (7.8% of the represented taxa) are invalid and 33 are valid but not updated (old designations or subspecies names missing). Third, only 75 strains (32.3%) are the type strains of the respective species, confirming previous observations[2,3]. The last point is really important, as a single strain is not representative of a species[4] but only the type strain is permanently linked to the name of a species[5,6].

Moreover, we found only 13 examples where genome typing with DNA microarrays was used to investigate the diversity of bacterial species and the type strain was included in the analysis. This dichotomy is depriving scientists of a unique framework for exchange and storage of information. For example, a large amount of data is accumulating on *Streptomyces coelicolor*, based on the genome sequence of *Streptomyces coelicolor* A3(2); however, strain A3(2) seems to be more closely related to *Streptomyces violaceoruber* than to the type strain of *Streptomyces coelicolor* (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=100226). Thus, in similar situations, the comprehension of the biology of a model organism and the extension of properties to similar organisms is impossible based on the name. Implications are broader, as names are crucial also for regulations in biotechnology and biosafety.

In the case of bacterial nomenclature, a standard already exists[5,6], but is not always followed. We therefore urge the following actions: first, authors should carefully analyze the taxonomic position of sequenced strains and evaluate and publish their relationships with the type strains by genome typing with DNA microarrays; second, journals and databases should apply strict policies concerning the taxonomic characterization and nomenclature of sequenced strains; and third, public sequencing programs should initiate the sequencing of type strains of important species to recover the link between genomics and the standard of bacterial nomenclature.

Moreover, the taxonomy of some bacterial groups (e.g., cyanobacteria) is not well defined and an effort is underway to improve the taxonomic schemes for bacterial biodiversity so that nomenclature rules can be applied broadly and consistently: to cite Bull, Ward and Goodfellow[7], "taxonomy is not a luxury."

Finally, taxonomy is a fast evolving area and the analysis of biodiversity often leads to the description of novel species and to nomenclatural changes with time. This means that names used in the papers can become obsolete, even if correct at the time of publication. This underscores the importance of correct and updated information in the online resources to put an end to "taxonomic anarchy"[8].

Note: Supplementary information is available on the Nature Biotechnology website.

COMPETING INTERESTS STATEMENT
The authors declare no competing financial interests.

Giovanna E Felis[1,2], Douwe Molenaar[1,3], Franco Dellaglio[2] & Johan E T van Hylckama Vlieg[1,3]

[1]NIZO Food Research, Kluyver Centre for Genomics of Industrial Fermentation, PO Box 20, 6710 BA Ede, The Netherlands. [2]Science and Technology Department, University of Verona, Strada le Grazie 15, I-37134 Verona, Italy. [3]Top Institute Food and Nutrition, PO Box