The Metadata Coverage Index (MCI): A standardized metric for quantifying database annotation richness

Konstantinos Liolios[1], Lynn Schriml[2], Lynette Hirschman[3], Ioanna Pagani[1], Bahador, Nosrat[1], Philippe Rocca-Serra[4], Susanna-Assunta Sansone[4], Chris Taylor[5], Nikos C. Kyrpides and Dawn Field [3,5]

[1] *Microbial Genomics and Metagenomic Super Program, Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, USA*
[2] *Department of Epidemiology and Public Health, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA.*
[3] *The MITRE Corporation, 202 Burlington Rd, Bedford MA 01730, USA*
[4] *Centre for Ecology & Hydrology, Maclean Building, Benson Lane, Crowmarsh Gifford, Wallingford, Oxfordshire, OX10 8BB*
[5] *European Molecular Biology Laboratory (EMBL) Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK*
[6] *University of Oxford, Oxford e-Research Centre, Oxford, OX1 3QG, UK.*

Abstract

Variability in the extent of the descriptions of data (metadata) held in public repositories forces users to assess the quality of records individually, which rapidly becomes impractical. The automatic scoring of records on the richness of their description enables sorting by quality. Here, we introduce an objective measure for metadata — the 'Metadata Coverage Index' (MCI): the percentage of available fields actually filled in a record or description. MCI scores can be calculated for a whole database, for individual records or for their component parts (variables or subsets of the data). The MCI score can be used to filter, rank or search for records, to assess the metadata quality of an *ad hoc* collection, or to determine the frequency with which fields in a particular record type are filled. Here we demonstrate the utility of MCI scores using metadata from the Genomes Online Database (GOLD), including records compliant with the 'Minimum Information about a Genome Sequence' standard developed by the Genomic Standards Consortium. Finally, we discuss a number of challenges and the further application of MCI score data to show improvements in annotation quality over time, to inform the work of standards bodies and repository providers on the usability and popularity of the same standards, and to credit the work of curators. Such an index provides a step towards putting metadata capture practices and in the future, standards compliance, into a quantitative and objective framework.

Introduction

"If you cannot measure it, you cannot improve it."
Lord Kelvin

As the size, number and complexity of bioscience data sets in the public domain continue to grow, appropriate contextualizing information becomes indispensible. Such 'halos' of information are referred to as metadata and include information on: how data were collected, processed and analyzed, the nature and state of the biological sample used and the research

context. Nowhere is this more relevant than in high-throughput studies using new technologies[1] , where the rate of production of data sets is becoming almost unmanageable given current public provision.

Metadata considered critical to data interpretation are often referred to as 'minimum information' (MI) and this concept has been expressed in various 'MI checklists'[2] covering a range of data types including transcriptomics, proteomics, metabolomics and genomics. MI checklists specify the contextual information that should be reported to ensure that studies are (in principle) reproducible and can be compared or combined in an appropriately-informed manner in downstream analyses. The increasing number of such specifications behooves the data-sharing community to develop methods to quantify the degree of compliance of databases, individual records or *ad hoc* collections, in order to highlight challenging-to-acquire components of specifications or to quantify improvements in metadata reporting or database content (for example, through curation).

Here we introduce a simple metric for evaluating the 'richness' of a database's metadata (or compliance with a given standard) and a straightforward method to calculate it. The 'Metadata Coverage Index' (MCI) is the number of fields in a record for which information is provided, expressed as a percentage of the total fields available. While this is an oversimplification (for reasons discussed below), it provides a starting point for quantification of the richness of information about an entry.

An MCI score represents arbitrarily complex contextual information as a simple numerical value. MCI scores can be calculated for individual fields or across collections/databases. While it is clear that some types of metadata carry more value than others, we have made no attempt to model distributions of value across database schemata or MI specifications. Establishment of a weighting among fields would be challenging and would be dependent on user requirements, but could be the focus of future work, along with the development of derived versions of MCI that weight particular types of information (i.e. depend on extended validation rules).

To illustrate the calculation of this metric and the usefulness of this concept, we use the MCI to examine the content of the Genomes Online Database (GOLD)[3] and assess compliance with the 'Minimum Information about a Genome Sequence' (MIGS) checklist[4] — a part of the MIxS standard [5] from the Genomic Standards Consortium (GSC)[6].

Materials and Methods

Data sets

Spreadsheets containing information for genomes from the Genomic Encyclopedia of Bacteria and Archaea (GEBA) and the Human Microbiome Project (HMP)[8] studies, as well as all the genome projects available from GOLD[3] were obtained from the GOLD database .

Calculation of MCI scores with the MCI Calculator

MCI scores were calculated for each collection as the total number of 'non-missing' fields expressed as a percentage of the total fields available across all records. Scores were also calculated for individual records and for each field (*i.e.*, each variable or column header in a spreadsheet). Note that MCI scores are expressed as percentages, and are therefore size-independent. While the scores could have been calculated using a spreadsheet, an MCI

Calculator tool was built to automate the process (Figure 1). As input, it takes any spreadsheet in tabular format. As output, MCI scores are calculated for the whole collection and new spreadsheets are generated containing the per-record and per-field scores. The MCI Calculator can be downloaded from http://genomesonline.org/SetupMCICalculator.msi .

For users: addition of MCI scores to the GOLD database
MCI scores were calculated for all records in GOLD and added to the GOLDCARD pages and the GOLD search interface. Therefore, MCI scores can now be used to search and sort GOLD records, for example, to see only records scoring above a certain threshold.

Results

Calculating MCI scores and comparison of metadata fields

The GOLD database, contains more than 200 metadata fields across more than 13,000 records, thus extending to well over 2.6 million data points[3]. For the purpose of this study, 113 metadata fields were selected from GOLD based on the fact that most of them apply for most types of projects, and their MCI scores were calculated across all genome records in the database as presented in Table 1.

There are five fields with MCI score 100 (fields 1-5 on Table 1). These are the fields that are filled with data across all the genome projects in GOLD. These are essential fields for project registration in the GOLD database. There are 7 more fields that have MCI score over 99 (fields 6-13). These are also essential fields for project registration, implying that most likely the data are missing due to an error, and they should also be filled in. Some of the fields in the list may seem redundant (e.g. fields 6 with 14 or fields 10 with 13), but when the number of records associated with them is displayed, they make better sense. For example GOLD has implemented a field named GOLD Genus (field 6), in addition to the Genus information provided from NCBI's Taxonomy (field 14). The reason is that the Genus information is available more readily at the time of the project registration in GOLD than it usually is at NCBI's taxonomy. The same with information related to the Phylum of the organism. The MCI score for the field NCBI Project ID is 75%, which implies that 25% of the projects in GOLD are not registered yet at NCBI's BioProject collection. 42% of the projects have Host Name information, which reflects the size of genome projects that are associated with a specific Host organism. 74% of the projects in GOLD have an update date (field 24 on Table 1), suggesting that the majority of the projects have been revisited for curation at least once after they were created in the database.

Overall, approximately two thirds of the 113 selected GOLD fields have an MCI score below 50 (fields 33-113). The MCI score for all 113 fields is 34.6. Ten of those fields apply only to projects that are part of the HMP study, and were excluded from subsequent comparisons across different datasets. Twelve fields are part of the MIGS fields as recommended by the GSC[4] (highlighted fields on Table 1). The position of the MIGS fields in the overall list of the 113 fields from GOLD, points to the fact that these are not the most frequently filled metadata fields across all the projects. Only two of the MIGS fields are among the top 10 GOLD fields and only six are among the top 50. While the MIGS fields were not likely to be the most populated fields (i.e. data for Isolation site and Latitude/Longitude are frequently not available, even though these are among the most important metadata fields), their overall position in the list certainly suggests that a revision may be necessary.

MCI score comparison of data sets

One advantage of calculating MCI scores as a percentage is that they are size independent and therefore allow direct comparison across collections. An MCI score captures the proportion of total *possible fields* that are *filled in* (have values) but do not enable a value judgment on the absolute number of values *present* in a particular collection. For comparison, Table 2 shows the MCI scores, along with the total number of records and fields, the maximum number of fields for each collection and the total number of filled values per collection.

We have created 9 distinct project collections from GOLD (Project list column on Table 2) and organized them in 3 separate groups for comparison. This allows comparison of the richness of various slices of the full database. Each comparison is meaningful only within its own group. For example, the "GEBA" collection is comprised of 256 genome projects, all part of the GEBA study. The collection "Complete" refers to the 2040 complete genome projects available in GOLD, while the "HMP" collection to the 2096 projects selected for sequencing under the HMP study. The collection "All projects" represents all the currently available 13790 isolate genome projects in GOLD, while the "Archaea", "Bacteria" and Eukarya" are the corresponding phylogenetic subgroups. Each project collection group is characterized by the specific number and type of fields selected for the comparison. This is essential in order to select fields that would be applicable for all the projects within a list. Accordingly, all the HMP related fields were excluded from the total number of fields used in this study, thus creating a set of 103 fields that apply to all project lists (CORE group). In a similar manner, the 10 HMP specific fields have been grouped to compose the HMP group, while the 12 MIGS fields composed the MIGS group of fields (all shown on the column Field group on Table 2).

Comparing the GEBA collection against the complete genomes, the HMP and the all projects lists, using the core 103 metadata fields (group A on Table 2), reveals that GEBA has the most well curated metadata projects, based on the highest MCI score (54.18%). This reflects the emphasis given to the collection and curation of metadata for this project. In terms of metadata coverage across different phylogenetic groups within the GOLD dataset (group 2, on Table 2), archaeal and bacterial subsets of the data had higher MCI scores than eukaryotes, pointing to the increased value of more detailed curation of the microbial genome projects in GOLD. Likewise, subsets of data compliant with the MIGS standard fields had relatively higher MCI scores, with the GEBA list reaching 68% of metadata coverage (group C on Table 2), almost 10% more than the average complete genome. Finally, within the HMP project list, the HMP fields have a high 70% MCI score (group D on Table 2).

Improvements in MCI scores over time

MCI scores can be used to compare collections and to quantify incremental increases in the richness of metadata over time. To illustrate this we compared the information contained in the GOLD database in 2008[9], 2010[10] and in 2012. The 2008 publication of GOLD reported a list of 45 metadata fields and the number of projects associated with those fields[9], while the 2010 publication of GOLD reported 105 variables and the number of projects for which information was available[10]. We selected a common set of 33 fields across the three sets and calculated the MCI scores for those (group E on Table 2). The results of this comparison

revealed that the overall MCI score has remained stable around 60%, although the total number of records has been doubling every two years. . These figures suggest that later submitters have tended to report more metadata, perhaps indicating greater acceptance of the need to provide appropriate metadata for submission.

Calculating MCI Scores for Records and Fields

MCI scores can be calculated for individual records or fields (variables) in a given dataset. This allows variation in MCI scores to be used to compare, sort and search records within datasets, or to select subsets. To show the utility of calculating MCI scores per record, MCI scores were included in the GOLD database. Using the advanced search option, users can now select records based on MCI score. For example, Figure 2 shows all entries with MCI scores > 50 on a world map, using associated metadata on the country of origin. The first ten projects in GOLD with the top MCI scores are shown on Table 3. Interestingly enough, six of those are part of the HMP study, showing that although the entire list of 2096 HMP projects has a relative low MCI score (39.91%), nevertheless some of the most well curated projects belong to this group. The remaining four projects are part of the Root Nodulating Bacteria (RNB) study running at the DOE Joint Genome Institute.

Discussion

We describe a new metric for assigning a numerical value to the richness of metadata in a given database. High MCI scores can indicate the most commonly-filled fields in existing records and so could be used to automatically select the most useful fields for display in tables or web interfaces (*i.e.*, the richest or most-commonly complete subsets of the data), or to empirically validate the content of a 'minimum information' specification[2]. The fields most frequently filled in a given collection are good candidates to be defined by a community as 'core'. If there is a mismatch – for example, if fields marked as 'core' in a standard are too difficult to collect, or fields with 100% compliance are not included in the standard – it suggests the content of that standard might need to be revised. This is particularly important given the recent GSC effort to define new habitat specific metadata fields (environmental packages) [5].

MCI scores, as calculated here, only take into account simple presence or absence of values. These scores could be further refined to take into account only valid entries, for example, those matching certain criteria (e.g., string, number, regular expression-compliant or curated versus calculated values), or those coming from recognized ontologies. This would be particularly useful for judging compliance with a given standard like MIGS, since free-text is not allowed and formal validation could be done, for example, using GCDML[12] (genomics) or the ISA-Tab (multi-omics) formats[13]. MCI scores could also be broken down to cover 'required' and 'optional fields' separately.

Refinement of MCI scores would depend on true validation of metadata which requires methods that encompass minimal information requirements, recommended terminologies, and also takes in account of the formats used to represent the metadata information. New efforts emerging from the community are laying the basis for a multi-dimensional validation process. Data commoning efforts, such as the ISA Commons[14], have delivered common

metadata tracking frameworks that can better underpin and facilitate the development of an improved validation method.

In the case of databases such as PRIDE[15] that allow unrestricted use of controlled vocabularies to extend records (i.e., user-defined fields), the list of available fields may appear disproportionately large (essentially each term used becomes a field, making for a very sparse matrix). MCI might not be appropriate for such data structures, but could be useful in helping to define whether a core (minimum) set of metadata can be defined for new data sets. The core would be the subset of the data with a very high relative MCI score.

When calculating MCI scores, it is important to consider that databases may also contain subsets where whole blocks of a record may be inapplicable; appropriate partitioning of records before score calculation would address this. Lastly, there are, of course, many databases where all fields are required; in these cases, there is no utility to scoring individual records.

In summary, we hope additional databases will adopt the use of MCI scores (and their derivatives) to highlight richness of associated metadata. MCI scores provide a solid step towards making possible the comparison of databases, *ad hoc* collections and individual records according to the richness of their metadata and of their component fields, providing valuable insights into the provenance, value and cost of the data so described. As such it is a metric that will help to place metadata capture practices and outputs into an objective and quantifiable framework.

Acknowledgements

This work was funded by NERC grant NE/D01252X/1 to DF. KL, IP, BN and NCK were supported by the Office of Science of the US Department of Energy under contract DE-AC02-05CH11231 and by the US National Institutes of Health Data Analysis and Coordination Center contract U01-HG004866. The support of Ioanna Bozionelou is especially acknowledged.

References

1 Field, D. *et al.* 'Omics Data Sharing. *Science* **326**, 234-236 (2009).

2 Taylor, C. F. *et al.* Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol* **26**, 889-896 (2008).

3 Pagani, I. *et al.* The Genomes On Line Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata *Nucleic Acids Res* **40**, D571-579, doi: 10.1093/nar/gkr1100 (2012).

4 Field, D. *et al.* The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* **26**, 541-547 (2008).

5 Yilmaz, P. *et al.* The "Minimum Information about a MARKer gene Sequence" (MIMARKS) specification. *Nat Biotechnol* **In press** (2010).

6 Field. The Genomic Standards Consortium (GSC). *PloS Biology* **(in press)** (2011).

7 Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**, 1056-1060, doi:nature08656 [pii] 10.1038/nature08656 (2009).

8 Peterson, J. *et al.* The NIH Human Microbiome Project. *Genome Res* **19**, 2317-2323 (2009).

9 Liolios, K. *et al.* The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **38**, D346-354, doi:gkp848 [pii] 10.1093/nar/gkp848 (2010).

10 Liolios, K. *et al.* The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **36**, D475-479, doi: 10.1093/nar/gkm884 (2008)

11 *GOLD - Genomes OnLine Database, Available at: http://www.genomesonline.org/*.

12 Kottmann, R. *et al.* A standard MIGS/MIMS compliant XML Schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *Omics : a journal of integrative biology* **12**, 115-121, doi:10.1089/omi.2008.0A10 (2008).

13 Rocca-Serra, P. *et al.* ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* **26**, 2354-2356, doi:btq415 [pii] 10.1093/bioinformatics/btq415 (2010).

14 Sansone, S. A. *et al.* Toward interoperable bioscience data. *Nat Genet* **44**, 121-126, doi:10.1038/ng.1054 (2012).

15 Jones, P. *et al.* PRIDE: new developments and new datasets. *Nucleic Acids Res* **36**, D878-883, doi:10.1093/nar/gkm1021 (2008).

Table 1. The list of all selected metadata fields in GOLD (columns 2 and 6), with the number of records for each of them (columns 3 and 7), and the MCI % (columns 4 and 8), ordered by the field with highest MCI.

	GOLD Metadata Field	Records	MCI %		GOLD Metadata Field	Records	MCI %
1	GOLD STAMP ID	13786	100	58	HMP FINISHING GOAL[1]	2472	17.93
2	DISPLAY NAME	13786	100	59	ENERGY SOURCES	2467	17.89
3	NCBI TAXON ID	13786	100	60	ASSEMBLY METHOD	2235	16.21
4	DOMAIN	13786	100	61	HMP ISOLATION BODY SITE[1]	2169	15.73
5	AVAILABILITY	13786	100	62	GREENGENES ID	2146	15.57
6	GOLD GENUS	13785	99.99	63	PROJECT DESCRIPTION	2122	15.39
7	PROJECT TYPE	13784	99.99	64	PUBLICATION LINK	2062	14.96
8	PROJECT STATUS	13784	99.99	65	HMP NCBI SUBMISSION STATUS[1]	1948	14.13
9	NCBI SUPERKINGDOM	13782	99.97	66	HMP PROJECT STATUS[1]	1948	14.13
10	GOLD PHYLUM	13778	99.94	67	HMP ID[1]	1946	14.12
11	PROPOSAL NAME	13761	99.82	68	ISOLATION SOURCE	1884	13.67
12	GOLD SPECIES	13734	99.62	69	SEQUENCING STATUS LINK	1849	13.41
13	NCBI PHYLUM	13526	98.11	70	GENE CALLING METHOD	1811	13.14
14	NCBI GENUS	13506	97.97	71	LONGITUDE	1631	11.83
15	NCBI ORDER	13435	97.45	72	LATITUDE	1629	11.82
16	NCBI SPECIES	13359	96.90	73	HMP ISOLATE SOURCE[1]	1482	10.75
17	NCBI FAMILY	13135	95.28	74	BEI STATUS[1]	1355	9.83
18	NCBI CLASS	13063	94.76	75	BODY SAMPLE SUBSITES	1236	8.97
19	SEQUENCING STATUS	12498	90.66	76	16S ID	1195	8.67
20	STRAIN	12480	90.53	77	BIOSAFETY LEVEL	1154	8.37
21	SEQUENCING COUNTRY	12326	89.41	78	ISOLATION DATE	1080	7.83
22	SEQUENCING CENTER	11837	85.86	79	HMP ISOLATION COMMENTS[1]	1052	7.63
23	NCBI PROJECT ID	10358	75.13	80	NUMBER OF READS	1048	7.60
24	UPDATE DATE	10247	74.33	81	ORGANISM COMMENTS	948	6.88
25	RELEVANCE	9993	72.49	82	METABOLISM	947	6.87
26	CONTACT NAME	8413	61.03	83	ISOLATION COMMENTS	874	6.34
27	HABITATS	7979	57.88	84	LIBRARY METHOD	778	5.64
28	TEMPERATURE RANGE	7673	55.66	85	SEROVAR	774	5.61
29	GRAM STAIN	7341	53.25	86	BODY PRODUCTS	723	5.24
30	BIOTIC RELATIONSHIP	7147	51.84	87	HOST HEALTH	712	5.16
31	CONTACT EMAIL	7037	51.04	88	STRAIN INFO ID	691	5.01
32	OXYGEN REQUIREMENT	7028	50.98	89	HMP ISOLATION COMMENTS[1]	690	5.01
33	CELL SHAPE	6748	48.95	90	HMP ISOLATION BODY SUBSITE[1]	681	4.94
34	DISEASES	6661	48.32	91	SYMBIOTIC RELATIONSHIP	493	3.58
35	MOTILITY	6275	45.52	92	SHORT READ ARCHIVE ID	475	3.45
36	HOST NAME	5807	42.12	93	INFORMATION URL	465	3.37
37	SEQUENCING METHODS	5636	40.88	94	PH	441	3.20
38	ISOLATION SITE	5388	39.08	95	IMAGE URL	415	3.01
39	SPORULATION	5187	37.63	96	VECTOR	380	2.76
40	HOST TAXON ID	5131	37.22	97	SYMBIONT	348	2.52
41	GENOME SIZE	4706	34.14	98	SYMBIOTIC INTERACTION	344	2.50
42	COMPLETION DATE	4585	33.26	99	ISOLATION PUBMED ID	339	2.46
43	CULTURE COLLECTION	4212	30.55	100	HOST GENDER	323	2.34
44	CELL ARRANGEMENTS	4126	29.93	101	DEPTH	308	2.23
45	PHENOTYPES	4045	29.34	102	SALINITY	281	2.04
46	GC PERC	3693	26.79	103	HOST AGE	250	1.81
47	GENE COUNT	3556	25.79	104	ISOLATION METHOD	238	1.73
48	IN IMG DATABASE	3453	25.05	105	CELL DIAMETER	233	1.69
49	PUBLICATION JOURNAL	3395	24.63	106	CELL LENGTH	189	1.37
50	SEQUENCING QUALITY	3286	23.84	107	COLOR	157	1.14
51	GEO LOCATION	3265	23.68	108	ALTITUDE	94	0.68
52	TYPE STRAIN	3248	23.56	109	HOST RACE	72	0.52
53	COVERAGE	3246	23.55	110	HOST COMMENTS	50	0.36
54	BODY SAMPLE SITES	3225	23.39	111	PROJECT COMMENTS	38	0.28
55	ISOLATION COUNTRY	3140	22.78	112	SYMBIONT TAXON ID	36	0.26
56	TEMPERATURE OPTIMUM	2712	19.67	113	NCBI ARCHIVE ID	10	0.07
57	CONTIG COUNT	2472	17.93				

[1] fields relevant only to projects that are part of the HMP study

Table 2. Comparison of MCI scores from the GOLD database. Note that if all variables in a database or collection apply to all records, then 'total fields' is equal to records multiplied by fields. If some variables are specific to a subset of records then the total number of possible fields will be smaller.

	Project List	Field group	Fields per Record	Records	Total Fields	Filled Fields	MCI %
A.	**1. GEBA**	CORE	103	256	26368	14287	54.18
	2. Complete			2040	211253	109532	52.00
	3. HMP			2096	215888	87007	39.91
	4. All Projects			13790	1420370	522850	37.00
B.	**1. Archaea**	CORE	103	340	35020	16767	48.00
	2. Bacteria			11233	1156999	443474	38.00
	3. Eukarya			2217	228351	62609	27.00
C.	**1. GEBA**	MIGS	12	256	3072	2102	68.43
	2. Complete			2040	24612	14667	59.59
	3. HMP			2096	25152	9642	38.34
	4. All Projects			13790	165480	62564	37.81
D.	**1. HMP**	HMP	10	2096	20960	14673	70.00
E.	**1. All-2008**	2008	33	2905	95865	59097	61.65
	2. All-2010			5843	192819	119881	62.17
	3. All-2012			13790	455070	273805	60.17

Table 3. The list of the genome projects in GOLD with the top 10 MCI scores

GOLD ID	Organism Name	Study Group	MCI %
Gi05215	Streptococcus bovis ATCC 700338	HMP	66.95
Gi02825	Mycobacterium parascrofulaceum ATCC BAA-614	HMP	66.10
Gc00590	Ensifer medicae WSM419	RNB	65.25
Gc00870	Rhizobium leguminosarum bv. trifolii WSM2304	RNB	65.25
Gi02071	Anaerofustis stercorihominis DSM 17244	HMP	64.41
Gi02072	Anaerotruncus colihominis DSM 17241	HMP	64.41
Gi02680	Clostridium hiranonis TO-931, DSM 13275	HMP	64.41
Gi01716	Clostridium scindens ATCC 35704	HMP	64.41
Gc01039	Rhizobium leguminosarum bv. trifolii WSM1325	RNB	64.41
Gi02147	Bacteroides stercoris ATCC 43183	RNB	63.56

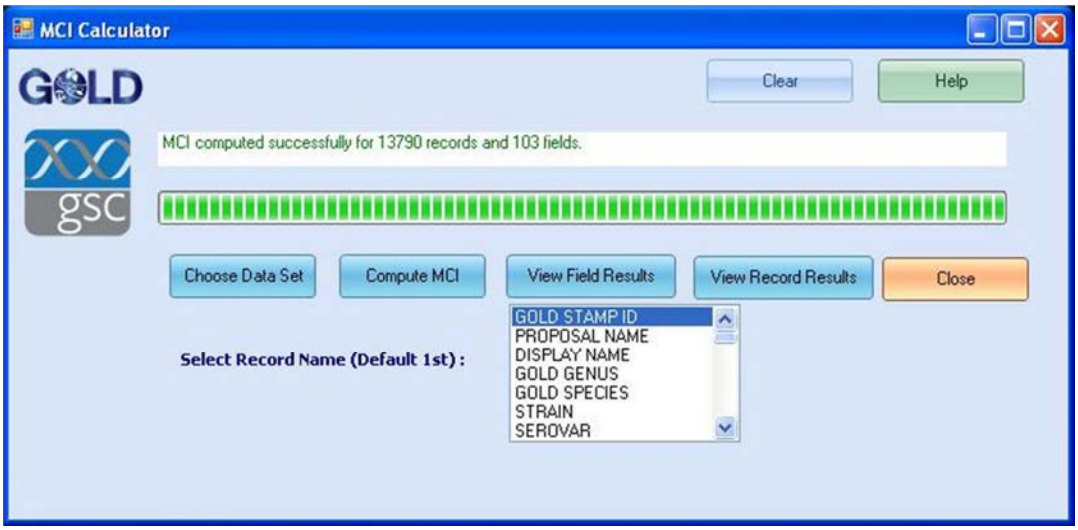

Figure 1. MCI calculator

Figure 2. MCI scores are implemented in the GOLD database. MCI scores can be seen on the GOLDCARDS for each entry and are including in the advanced search option. For example, all entries with an MCI score > 50 are shown on the map below.