EBI is an Outstation of the European Molecular Biology Laboratory.

Proteomics data sharing: why?

- Since the Enlightenment, Science is built upon the sharing of knowledge and information

  "*If I have looked further, it is because I stood on the shoulder of Giants*"
  -- Sir Isaac Newton, paraphrasing Bernard of Chartres

- Data producers are not always the best data analysts

  Sharing of data allows analysts access to real data, and in turn allows better analysis tools to be developed

- Meta-analysis of data can recycle previous findings for new tasks

  putting findings in the context of other findings increases their scope

- Sharing data allows independent review of the findings

  When actual replication of an experiment is often impossible, a re-analysis or spot checks on the obtained data become vitally important

- Simple economics

  "Information, no matter how expensive to create, can be replicated and shared at little or no cost." -- Thomas Jefferson

Simply sharing data is not enough…

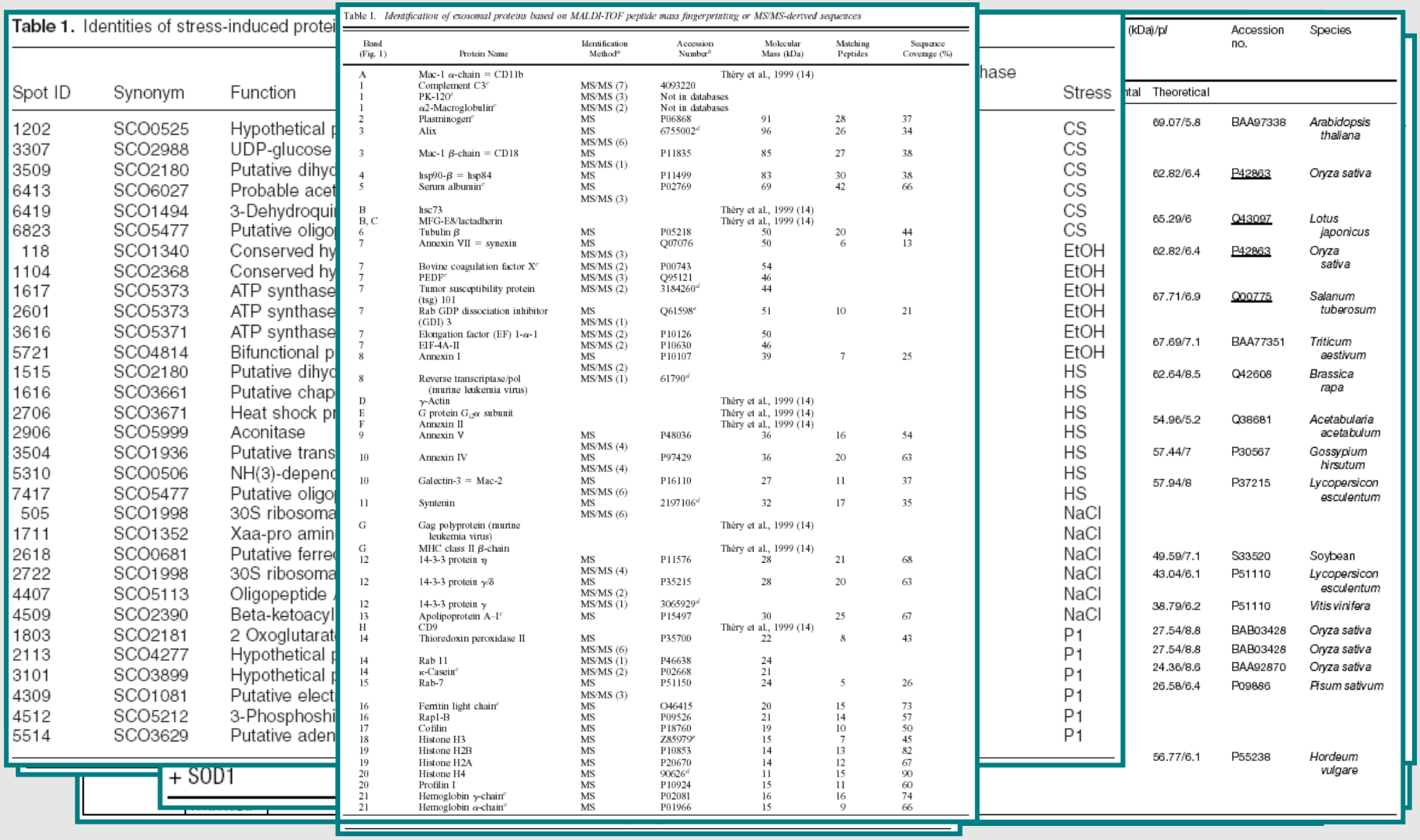


Table 1. Identities of stress-induced protei

Spot ID	Synonym	Function
1202	SCO0525	Hypothetical p
3307	SCO2988	UDP-glucose
3509	SCO2180	Putative dihyc
6413	SCO6027	Probable acet
6419	SCO1494	3-Dehydroquii
6823	SCO5477	Putative oligo
118	SCO1340	Conserved hy
1104	SCO2368	Conserved hy
1617	SCO5373	ATP synthase
2601	SCO5373	ATP synthase
3616	SCO5371	ATP synthase
5721	SCO4814	Bifunctional p
1515	SCO2180	Putative dihyc
1616	SCO3661	Putative chap
2706	SCO3671	Heat shock pr
2906	SCO5999	Aconitase
3504	SCO1936	Putative trans
5310	SCO0506	NH(3)-depend
7417	SCO5477	Putative oligo
505	SCO1998	30S ribosoma
1711	SCO1352	Xaa-pro amin
2618	SCO0681	Putative ferre
2722	SCO1998	30S ribosoma
4407	SCO5113	Oligopeptide
4509	SCO2390	Beta-ketoacyl
1803	SCO2181	2 Oxoglutarat
2113	SCO4277	Hypothetical p
3101	SCO3899	Hypothetical p
4309	SCO1081	Putative elect
4512	SCO5212	3-Phosphoshi
5514	SCO3629	Putative aden

+ SOD1

Table I. *Identification of exosomal proteins based on MALDI-TOF peptide mass fingerprinting or MS/MS-derived sequences*

Band (Fig. 1)	Protein Name	Identification Method[a]	Accession Number[b]	Molecular Mass (kDa)	Matching Peptides	Sequence Coverage (%)
A	Mac-1 α-chain = CD11b		Théry et al., 1999 (14)			
1	Complement C3[c]	MS/MS (7)	4093220			
1	PK-120[c]	MS/MS (3)	Not in databases			
1	α2-Macroglobulin[c]	MS/MS (2)	Not in databases			
2	Plasminogen[c]	MS	P06868	91	28	37
3	Alix	MS	6755002[d]	96	26	34
		MS/MS (6)				
3	Mac-1 β-chain = CD18	MS	P11835	85	27	38
		MS/MS (1)				
4	hsp90-β = hsp84	MS	P11499	83	30	38
5	Serum albumin[c]	MS	P02769	69	42	66
		MS/MS (3)				
B	hsc73		Théry et al., 1999 (14)			
B, C	MFG-E8/lactadherin		Théry et al., 1999 (14)			
6	Tubulin β	MS	P05218	50	20	44
7	Annexin VII = synexin	MS	Q07076	50	6	13
		MS/MS (3)				
7	Bovine coagulation factor X[c]	MS/MS (2)	P00743	54		
7	PEDF[c]	MS/MS (3)	Q95121	46		
7	Tumor susceptibility protein (tsg) 101	MS/MS (2)	3184260[d]	44		
7	Rab GDP dissociation inhibitor (GDI) 3	MS	Q61598[e]	51	10	21
		MS/MS (1)				
7	Elongation factor (EF) 1-α-1	MS/MS (2)	P10126	50		
7	EIF-4A-II	MS/MS (2)	P10630	46		
8	Annexin I	MS	P10107	39	7	25
		MS/MS (2)				
8	Reverse transcriptase/pol (murine leukemia virus)	MS/MS (1)	61790[d]			
D	γ-Actin		Théry et al., 1999 (14)			
E	G protein $G_{i2}\alpha$ subunit		Théry et al., 1999 (14)			
F	Annexin II		Théry et al., 1999 (14)			
9	Annexin V	MS	P48036	36	16	54
		MS/MS (4)				
10	Annexin IV	MS	P97429	36	20	63
		MS/MS (4)				
10	Galectin-3 = Mac-2	MS	P16110	27	11	37
		MS/MS (6)				
11	Syntenin	MS	2197106[d]	32	17	35
		MS/MS (6)				
G	Gag polyprotein (murine leukemia virus)		Théry et al., 1999 (14)			
G	MHC class II β-chain		Théry et al., 1999 (14)			
12	14-3-3 protein η	MS	P11576	28	21	68
		MS/MS (4)				
12	14-3-3 protein γ/δ	MS	P35215	28	20	63
		MS/MS (2)				
12	14-3-3 protein γ	MS/MS (1)	3065929[d]			
13	Apolipoprotein A–I[c]	MS	P15497	30	25	67
H	CD9		Théry et al., 1999 (14)			
14	Thioredoxin peroxidase II	MS	P35700	22	8	43
		MS/MS (6)				
14	Rab 11	MS/MS (1)	P46638	24		
14	κ-Casein[c]	MS/MS (2)	P02668	21		
15	Rab-7	MS	P51150	24	5	26
		MS/MS (3)				
16	Ferritin light chain[c]	MS	O46415	20	15	73
16	Rap1-B	MS	P09526	21	14	57
17	Cofilin	MS	P18760	19	10	50
18	Histone H3	MS	Z85979[e]	15	7	45
19	Histone H2B	MS	P10853	14	13	82
19	Histone H2A	MS	P20670	14	12	67
20	Histone H4	MS	90626[d]	11	15	90
20	Profilin I	MS	P10924	15	11	60
21	Hemoglobin γ-chain[c]	MS	P02081	16	16	74
21	Hemoglobin α-chain[c]	MS	P01966	15	9	66

hase

Stress
CS
CS
CS
CS
CS
CS
EtOH
EtOH
EtOH
EtOH
EtOH
EtOH
HS
HS
HS
HS
HS
HS
HS
NaCl
NaCl
NaCl
NaCl
NaCl
P1
P1
P1
P1
P1
P1

(kDa)/pI		Accession no.	Species
ital	Theoretical		
	69.07/5.8	BAA97338	*Arabidopsis thaliana*
	62.82/6.4	P42863	*Oryza sativa*
	65.29/6	Q43097	*Lotus japonicus*
	62.82/6.4	P42863	*Oryza sativa*
	67.71/6.9	Q00775	*Salanum tuberosum*
	67.69/7.1	BAA77351	*Triticum aestivum*
	62.64/8.5	Q42608	*Brassica rapa*
	54.96/5.2	Q38681	*Acetabularia acetabulum*
	57.44/7	P30567	*Gossypium hirsutum*
	57.94/8	P37215	*Lycopersicon esculentum*
	49.59/7.1	S33520	Soybean
	43.04/6.1	P51110	*Lycopersicon esculentum*
	38.79/6.2	P51110	*Vitis vinifera*
	27.54/8.8	BAB03428	*Oryza sativa*
	27.54/8.8	BAB03428	*Oryza sativa*
	24.36/8.6	BAA92870	*Oryza sativa*
	26.58/6.4	P09886	*Pisum sativum*
	56.77/6.1	P55238	*Hordeum vulgare*

EMBL-EBI

A nuance: *available* data vs. *accessible* data

When data is only made available as arbitrarily formatted PDF tables, it carries important limitations

- Source data (e.g., mass spectra) are not made available
  - No peer review validation possible (*the numbers game*)
  - Very little raw materials for testing innovative *in silico* techniques are available (*data hoarding*)
- Automated (re-)processing of the results (e.g., identifications) is impossible (*eliminating objective technique comparison*)
- Data producers do not actually feed their results and knowledge back to the community (*evading responsability for the results*)

Accessibility requires proper infrastructure

- Community supported, standardized data formats
  Necessary to allow efficient access to the data

- Controlled vocabularies (CV's) and ontologies
  To provide unambiguous context and metadata to the actual data, as well as enabling powerful queries to be performed on the data

- Minimal reporting requirements for specific data types
  Ensures the presence of certain bits of information without which interpretation is ambiguous, hampered or impossible

- Publicly available, online repositories
  Bioinformatics grew up along side the internet, and this is reflected in the successful online data sharing mechanims already in place in the life sciences. *The repositories should implement the standards, use the CV's and ontologies, and adhere to the minimal requirements.*

Community standards for proteomics

The Human Proteome Organisation (HUPO)
Proteomics Standards Initiative (PSI)

http://www.psidev.info

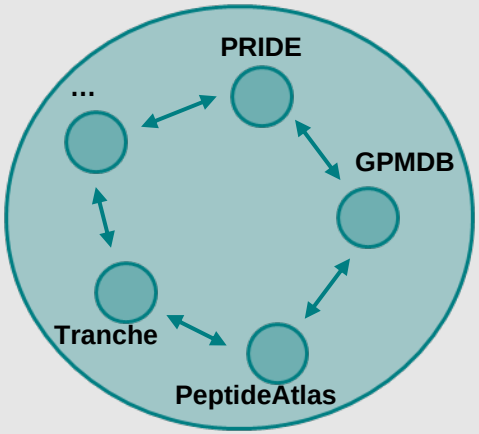
- **Creates standard formats, CV's and ontologies, and minimal requirements**
- **Composed of several workgroups**

Molecular Interactions	***(MI)***	***PSI-MI format v2.5***
Mass Spectrometry	***(MS)***	***mzData, mzML format***
Sample Processing	***(SP)***	***alpha stage***
Gel Electrophoresis	***(Gel)***	***GelML format***
Proteomics Informatics	***(PI)***	***analysisXML format***
Protein Modifications	***(Mod)***	***PSI-MOD ontology***

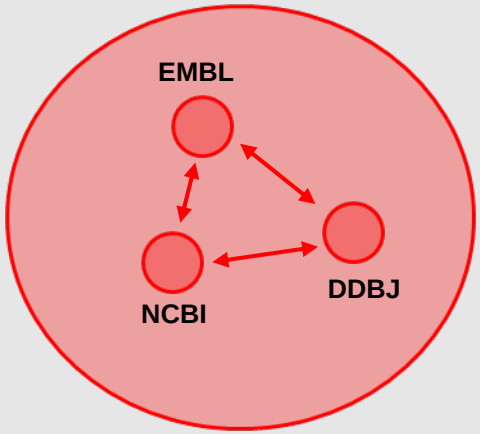
Furthermore, superstructures must be built

Often, multiple repositories will emerge more or less simultaneously in a particular field. By exchanging data, and by collaborating on data acquisition an increase in coverage as well as a more comprehensive dataset is obtained by each individual resource.

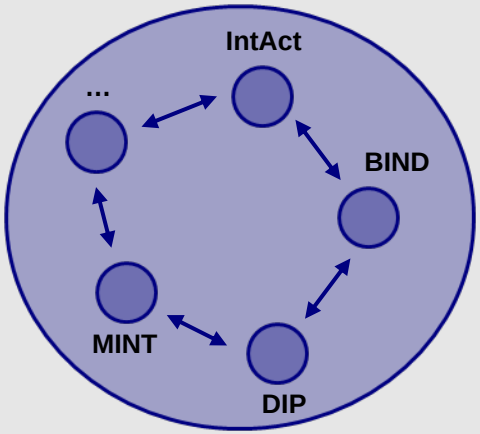
Such superstructures do require additional infrastructure, however.



mass spec
ProteomExchange



sequence databases
(INSDC)

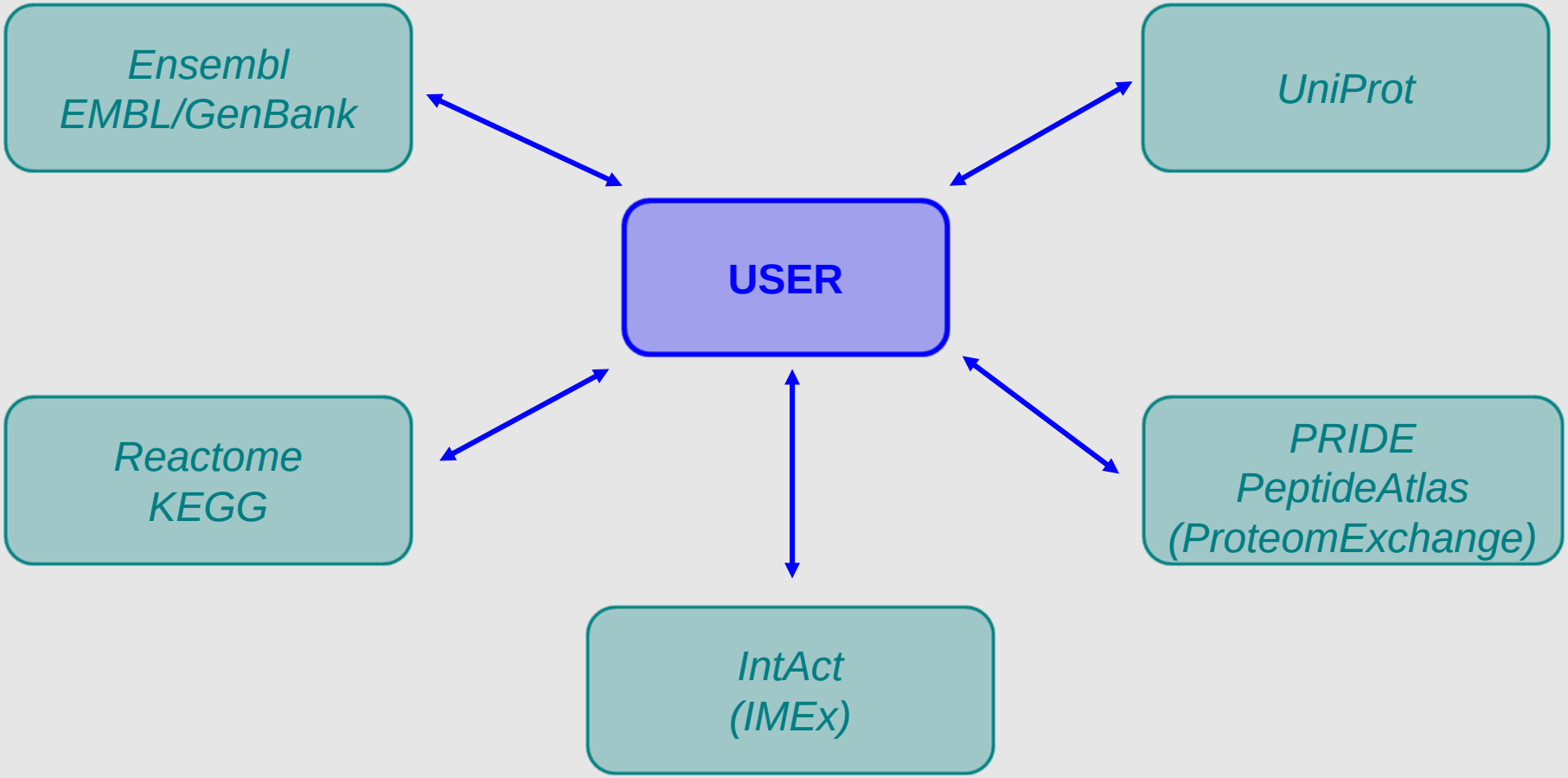


interactions
IMEx

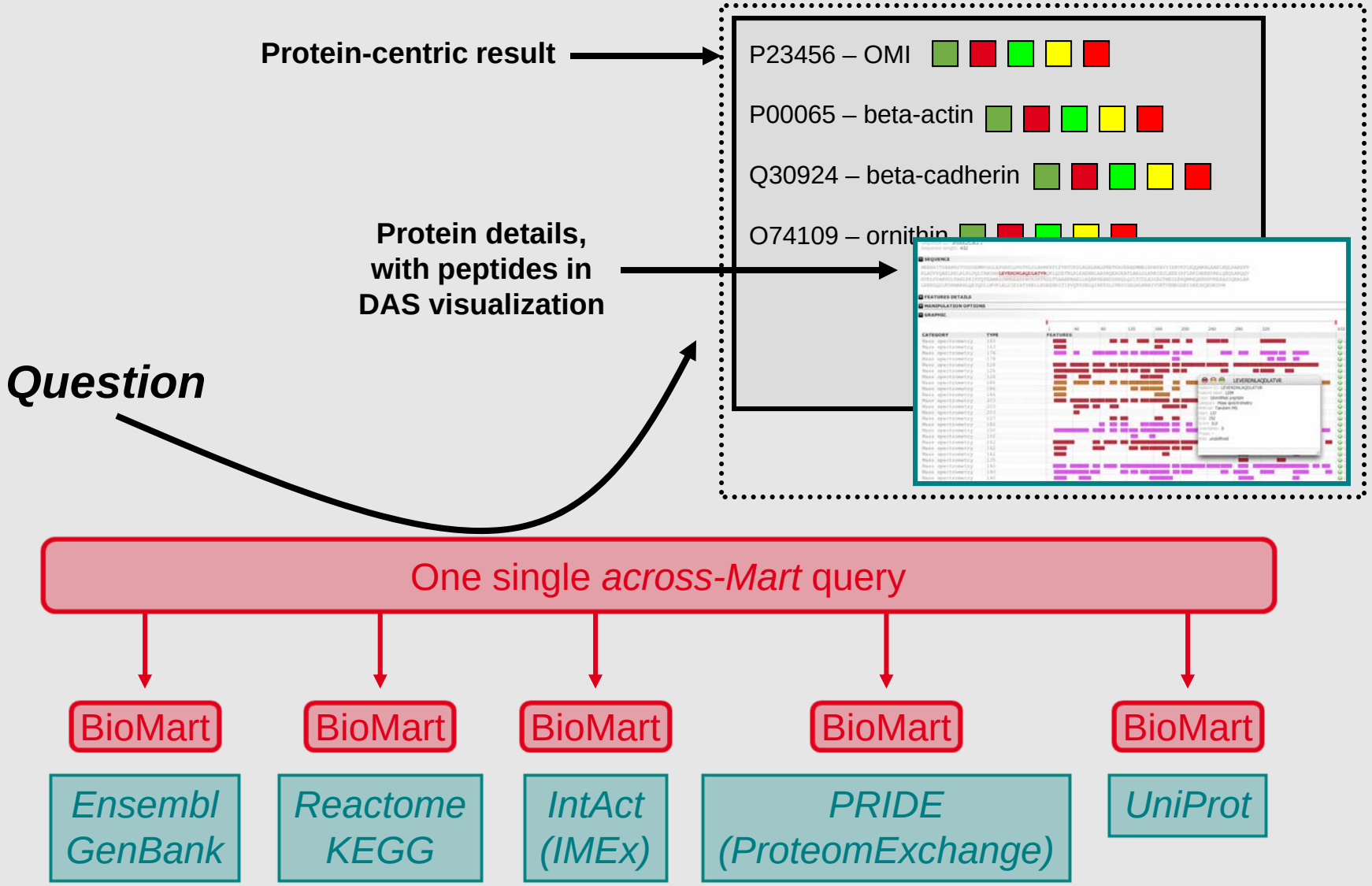
The final goal: cross-domain integration

Current situation

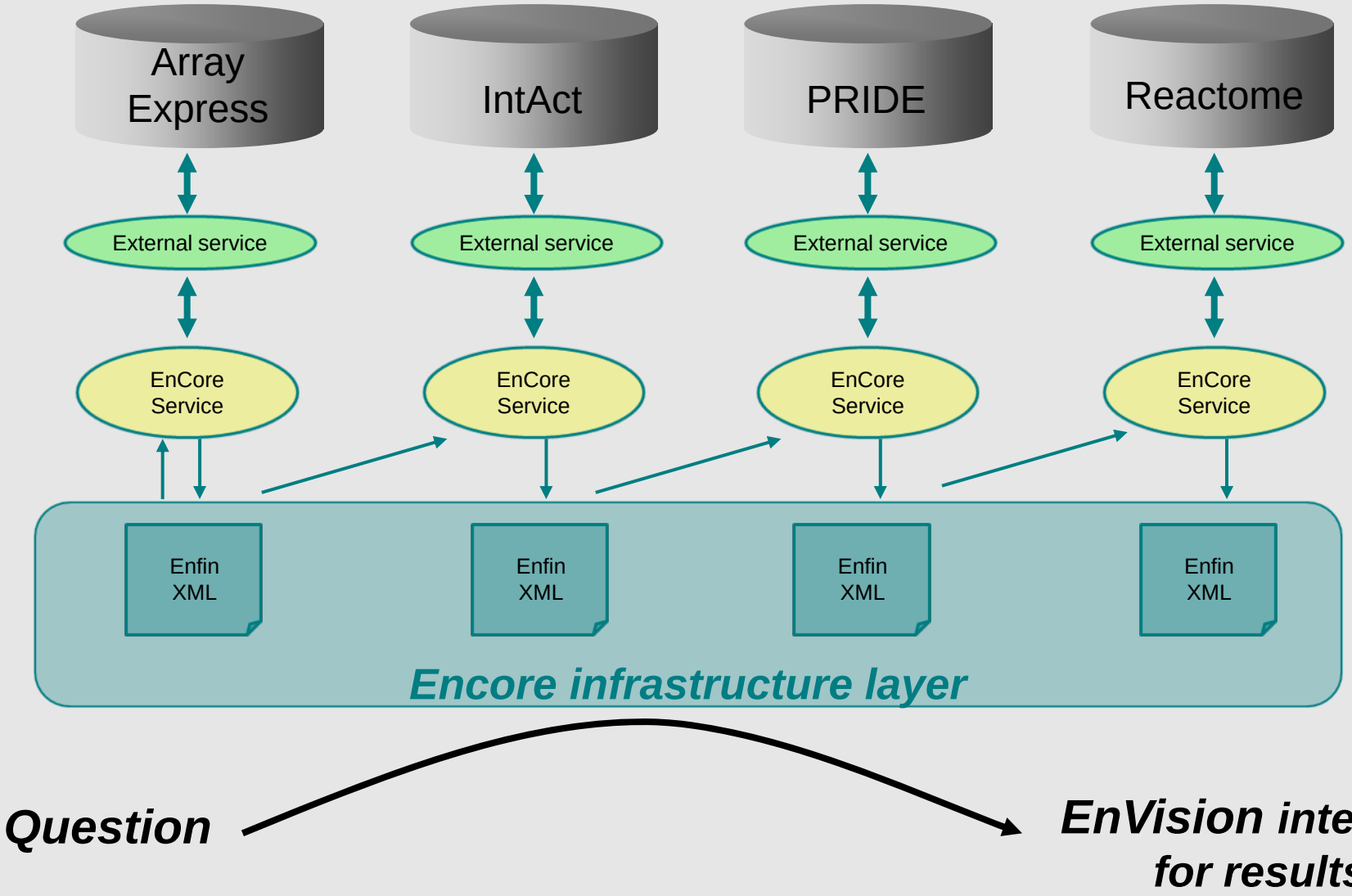
(requires user expertise in querying all the different repositories)



BioMart as a common interface layer



Enfin/Encore as a cross-domain aggregator



How do we make this all happen?

- Journal guidelines

  Journal guidelines heavily influence the decisions taken by authors; by first requesting and subsequently mandating data submission to established repositories, they provide an important stick.

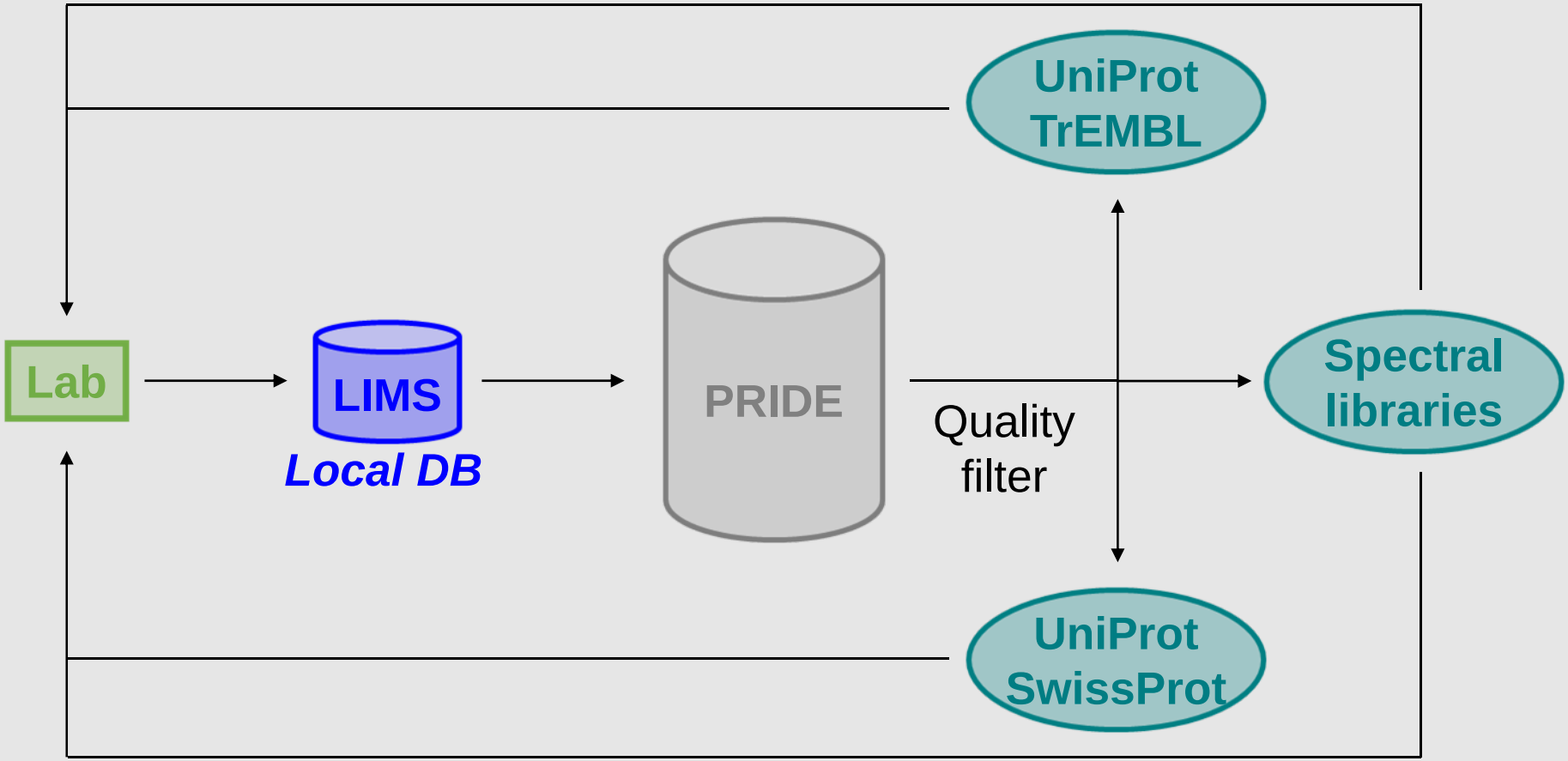
- Funder support and guidelines

  Funders contribute both sticks and carrots. The sticks lie in the grant application guidelines; they can require a plan for data management and dissemination. The carrot is in providing specific funding for this aspect of science.

- Data repositories

  The availability of reliable, freely available repositories is key; submission thresholds should be kept low and added value needs to be provided. Furthermore, feedback loops need to be established in order to ensure that accumulated data flows back to the user community. Repositories thus provide mostly carrots.

An example of a proteomics feedback loop



Funding example: EC FP6 ProDaC grant

- **Coordination Action grant in the 6th Framework Programme of the European Commission**
- **Good example of a coordination between: (i) funder, (ii) data producers, (iii) repository providers, (iv) standards organization, and (v) journals**
- **Workpackages**

WP1: standards for data representation	***PSI standards development***
WP2: Standards implementation	***provide compatible software***
WP3: Data integration tools	***enabling data submission***
WP4: Proteomics repository adaptation	***ensuring repository compliance***
WP5: Data flow management	***overall data gathering and submission***
WP6: Proteomics data exploitation	***establishing data feedback loops***
(WP7: project management	***overall project management)***