

La numérisation durable

(Sustainable digitization)*

Denis Roegel

Université de Lorraine & LORIA

21 décembre 2013

* (une version anglaise est aussi disponible en
<http://locomat.loria.fr/other/num.html>)

Préambule

Nombreuses questions

- faut-il numériser ?
- que numériser ?
- pourquoi ?
- comment ?
- qui doit numériser ?
- qu'est-ce que la **numérisation durable** ?

L'objectif de cette présentation est de fournir des **éléments de réponse** à ces questions.

Numérisation durable : concept nouveau ?

- Au 21^e siècle, la **durabilité** devient le mot phare.
- Mais curieusement, il a très peu été employé pour la numérisation. Pourquoi ?
- ce concept semble mal compris et récemment, un professeur de l'ENSSIB m'écrivait : « À ma connaissance, la numérisation durable n'existe pas ».
- est-ce vraiment le cas ?

Notre contexte

- nous avons été sensibilisé depuis près de vingt ans à deux facettes fondamentales de la qualité des documents : celle des **mathématiques** et celle des **objets graphiques** ;
- plus récemment, la fréquentation et l'utilisation intensive de documents numérisés nous a conduit à avoir une **approche critique** ;
- une réflexion sur les **besoins des utilisateurs**, les **besoins des documents**, et la **dynamique de la numérisation** et de ses acteurs a été dégagée ;
- des échanges avec les bibliothèques ont permis de mieux comprendre les **objectifs** et les **a priori** des uns et des autres ;
- notre réflexion a notamment été stimulée par notre travail de **reconstruction de documents historiques**, le projet LOCOMAT.

Un mot sur le projet LOCOMAT

LOCOMAT (<http://locomat.loria.fr>) est un site dédié aux tables mathématiques et astronomiques, avec actuellement deux aspects très développés :

- 1 une **reconstruction de tables** historiques (100000 pages)
 - 2 un **recensement de toutes les tables numérisées**, de toutes les époques, sur tous les supports et dans toutes les langues (près de 2000 ouvrages)
- le recensement de LOCOMAT fournit un aperçu sans doute représentatif de la numérisation des tables mathématiques et astronomiques
 - les fonds, interfaces et qualité de numérisation peuvent être comparés

Le problème de la qualité

Une grande partie de cette présentation tourne autour de la notion de **qualité**.

- un préalable pour la suite de la présentation est d'avoir pour objectif une certaine qualité ;
- **cette présentation ne s'adresse donc pas à ceux qui n'auraient que des intérêts commerciaux ou économiques ;**
- la qualité est liée à la durabilité : on ne veut pas refaire le travail déjà fait ;
- si la qualité et la durabilité vous intéressent, lisez ce qui suit, sinon, passez votre chemin.

Le contexte des archives (1)

Avant de commencer, il est nécessaire de dire un mot sur la numérisation dans les Archives. Il y a une différence fondamentale entre *archives* et *bibliothèques* :

- les bibliothèques conservent des livres qui ont naturellement été destinés à être diffusés ;
- les archives conservent des documents habituellement uniques, non destinés à être diffusés.

Le contexte des archives (2)

Cette différence fait que les archivistes ne se sentent pas concernés par ce qui se passe dans les bibliothèques. La numérisation sert pour eux essentiellement à deux besoins :

- nécessité de protéger des originaux
- besoins d'accès pour quelques types particuliers de documents (registres paroissiaux, actes d'état civil et actes notariés)

En outre, il faut noter que :

- la numérisation permet de désengorger les lieux d'archives (les originaux ne peuvent pas sortir) ;
- la numérisation concerne en grande majorité des manuscrits ;
- il n'y a pas du tout de vision de la numérisation à long terme et la notion de numérisation durable ne semble pas concerner les archivistes.

Le contexte des archives (3)

Mais, tout comme dans les bibliothèques, le processus de numérisation des archives se poursuivra :

- on choisira toujours des fonds à numériser parmi ceux encore non numérisés et, peu à peu se dessinera l'objectif de la numérisation totale ;
- même si les documents des archives n'ont pas été destinés initialement à la diffusion, celle-ci interviendra car les besoins apparaîtront ;

Au niveau des livres, la politique de numérisation semble plus claire, mais c'est la seule différence.

Le contexte des archives (4)

- les problèmes techniques sont en fait communs, même si certains archivistes pensent que la numérisation des livres, « c'est du gâteau » ;
- si les livres étaient vraiment si simples à numériser, on se demande pourquoi c'est presque toujours mal fait ;
- dans tous les cas, les archivistes doivent être conscients du problème de numérisation de documents peut-être plus simples que les leurs ;

Le contexte des archives (5)

- finalement, passé le cap des priorités, la numérisation durable devient une réalité, même si elle est naturellement très éloignée dans le temps (tout comme dans les bibliothèques elle a été *artificiellement* placée à court-terme) ;
- la presque totalité des analyses de notre document s'appliquent en fait aux archives, mais dans un contexte de durabilité.

Remarque importante :

Ce document n'est pas un manuel de numérisation
et seuls certains aspects pratiques et techniques sont évoqués.

Nous allons maintenant pouvoir commencer.

Fin du préambule

- 1 La qualité des numérisations
- 2 La qualité des interfaces
- 3 Les causes des problèmes techniques
- 4 Les problèmes de politique de numérisation
- 5 Solutions

La qualité des numérisations

- dans ce qui suit, nous allons faire un panorama d'un certain nombre de numérisations représentatives ;
- ces numérisations seront commentées et cela permettra de dégager des besoins ou des écueils à éviter.

Exemples de mauvaises numérisations

Andoyer : « La théorie de la lune » (1902)

Gallica (à partir d'un microfilm probablement presque aussi mauvais)

ii

THÉORIE DE LA LUNE

lité à longue période de la longitude, on a, par l'équation (11), d'après les propriétés de la fonction B :

$$\begin{aligned}
 & i(m - g_0)b_{17} + 2(1 - m)b_1a_{17}^2 + 2(1 - g_0)b_{39} \left(\frac{3}{8}m^2 \frac{2}{2(1 - m)} \right) \\
 & + (1 - g_0)b_{17} \left[2a_1a_{17} + \frac{3}{8}m^2(-2a_{17} - 2b_{17}) \frac{2}{1 - 2m - g_0} + a_{17} \frac{3}{8}m^2 \frac{2}{2(1 - m)} \right] \\
 & - ia_1a_{17}^2 - 6a_0a_{17}a_{19} + ii(m - g_0)^2a_{17}a_{19} \\
 & + 10m^2a_{17}a_{19} + \frac{3}{4}m^2[-7a_{39} - 8b_{39} + 10a_{17}^2 + 14a_{17}b_{17} + 8b_{17}^2] = 0.
 \end{aligned}$$

La formule en clair formatée avec un outil moderne

$$\begin{aligned}
 & 4(m - g_0)b_{41} + 2(1 - m)b_1a_{17}^2 + 2(1 - g_0)b_{39} \left(\frac{3}{8}m^2 \frac{2}{2(1 - m)} \right) \\
 & + (1 - g_0)b_{17} \left[2a_1a_{17} + \frac{3}{8}m^2(-2a_{17} - 2b_{17}) \frac{2}{1 - 2m - g_0} + a_{17} \frac{3}{8}m^2 \frac{2}{2(1 - m)} \right] \\
 & - 3a_1a_{17}^2 - 6a_0a_{17}a_{19} + 44(m - g_0)^2a_{17}a_{19} \\
 & + 10m^2a_{17}a_{19} + \frac{3}{4}m^2[-7a_{39} - 8b_{39} + 10a_{17}^2 + 14a_{17}b_{17} + 8b_{17}^2] = 0.
 \end{aligned}$$

Exemples de mauvaises numérisations

Andoyer : « La théorie de la lune » (1902)

Le même extrait
(inutilisable) de Gallica

ii

THÉORIE DE LA LUNE

lité à longue période de la longitude, on a, par l'équation (15),
d'après les propriétés de la fonction B :

$$\begin{aligned}
 & i(m - g_0) b_{11} + 2(1 - m) b_1 a_{17}^2 + 2(1 - g_0) b_{39} \left(\frac{3}{8} m^2 \frac{2}{1 - m} \right) \\
 & + (1 - g_0) b_{17} \left[2a_1 a_{17} + \frac{3}{8} m^2 (-2a_{17} - 2b_{17}) \frac{2}{1 - 2m - g_0} \right. \\
 & \quad \left. + a_{17} \frac{3}{8} m^2 \frac{2}{2(1 - m)} \right] \\
 & - 3a_1 a_{17}^2 - 6a_0 a_{17} a_{19} + 44(m - g_0)^2 a_{17} a_{19} \\
 & + 10m^3 a_{17} a_{19} + \frac{3}{4} m^2 [-7a_{39} - 8b_{39} + 10a_{17}^2 + 14a_{17} b_{17} + 8b_{17}^2] = 0.
 \end{aligned}$$

Le même passage en
scannant une photocopie
de l'original :

$$\begin{aligned}
 & 4(m - g_0) b_{11} + 2(1 - m) b_1 a_{17}^2 + 2(1 - g_0) b_{39} \left(\frac{3}{8} m^2 \frac{2}{1 - m} \right) \\
 & + (1 - g_0) b_{17} \left[2a_1 a_{17} + \frac{3}{8} m^2 (-2a_{17} - 2b_{17}) \frac{2}{1 - 2m - g_0} \right. \\
 & \quad \left. + a_{17} \frac{3}{8} m^2 \frac{2}{2(1 - m)} \right] \\
 & - 3a_1 a_{17}^2 - 6a_0 a_{17} a_{19} + 44(m - g_0)^2 a_{17} a_{19} \\
 & + 10m^3 a_{17} a_{19} + \frac{3}{4} m^2 [-7a_{39} - 8b_{39} + 10a_{17}^2 + 14a_{17} b_{17} + 8b_{17}^2] = 0.
 \end{aligned}$$

Exemples de mauvaises numérisations (suite)

- En 2009 ou 2010, la ville de Turin a numérisé un important ouvrage de Giovanni Plana sur la théorie de la lune, soit environ 2000 pages.
- Plana était originaire de Turin, et c'était une mise en valeur de son patrimoine
- <http://www.accademiadelle scienze.it/TecaRicerca>
- mais :
 - le site est difficile à utiliser car nécessite une configuration non universelle
 - la résolution est insuffisante
 - le téléchargement est impossible
- cette numérisation aura donc été un simple **gaspillage**
- entre-temps, l'ouvrage a été (un peu) mieux numérisé sur Internet Archive

Exemples de mauvaises numérisations (suite)

Plana : « Théorie du mouvement de la lune » (1832), vol. 2, p. 216
 (Version d'Archive.org) **exposants difficiles à lire**

$$(4) \dots \dots - R_1 \cdot \frac{ds_1}{dv} =$$

$$g\nu \quad e\gamma \left\{ \frac{3}{2} - \frac{3}{2} m + \frac{9}{8} m^2 + \frac{9}{8} e^2 - \frac{15}{4} \varepsilon^{1/2} - \left(\frac{3}{8} + \frac{3}{8} = \right. \right.$$

$$g\nu \quad e\gamma \left\{ \frac{3}{2} + \frac{3}{2} m + \frac{9}{8} m^2 + \frac{9}{8} e^2 - \frac{15}{4} \varepsilon^{1/2} - \left(\frac{3}{8} + \frac{3}{8} = \right. \right.$$

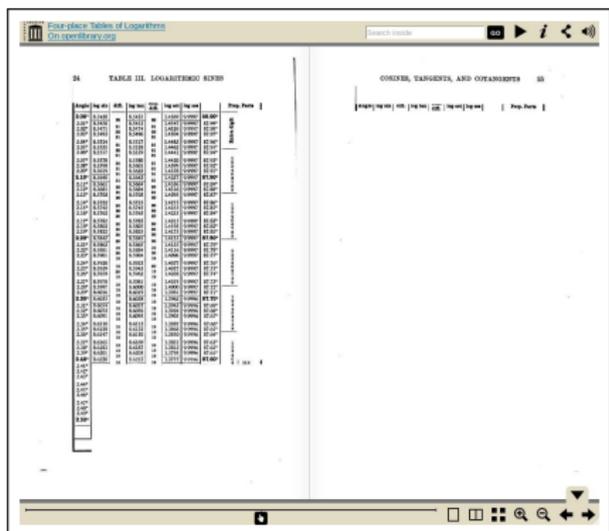
$$-g\nu \quad e^2\gamma \left\{ - \frac{15}{8} + \frac{57}{16} m \right\}$$

$$\left(\frac{15}{8} \frac{57}{16} \dots \dots / 2991 \quad 45 \quad 2271 \backslash \dots \dots \frac{15}{8} \dots \right)$$

(Archive et Google appliquent des seuils qui réduisent la qualité.)

Exemples de mauvaises numérisations (suite)

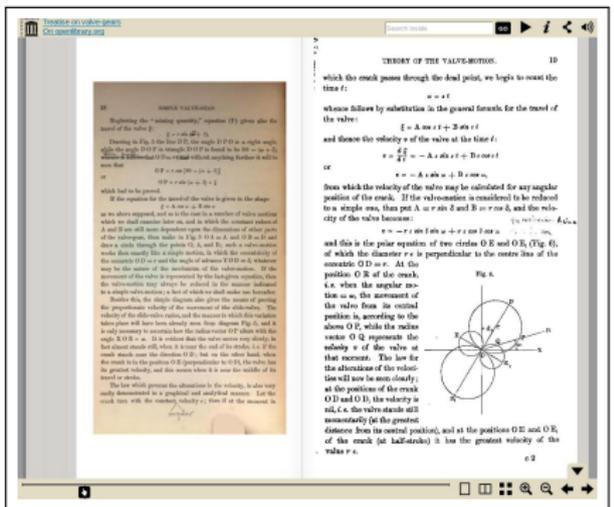
Comme Archive inclut des numérisations de Google, on ne s'étonnera pas que certaines pages soient totalement anéanties :



- Granville : *Four-place tables of logarithms*, 1908
- id = fourplacetable00grangoog

Exemples de mauvaises numérisations (suite)

Archive présente des numérisations incohérentes :



- Zeuner, *Treatise on valve-gears*, 1869
- id = [treatiseonvalve01zeungoog](#)
- pages gauches et droites différentes

Exemples de mauvaises numérisations (suite)

Felkel : « Tafel aller einfachen Faktoren » (1776)
 (<http://resolver.sub.uni-goettingen.de/purl?PPN620754826>)

Factoren von 3600 bis 42000.

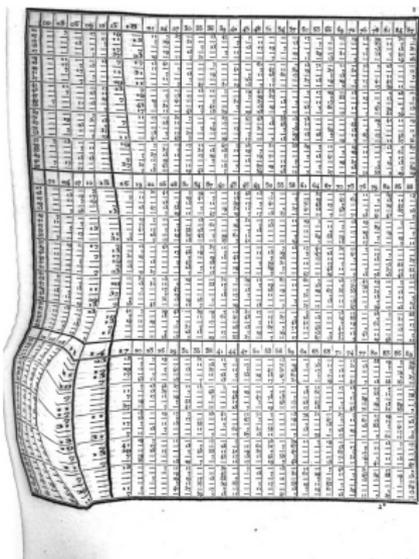
t	a	b	c	d	e	f	g	h	i	a	b	c	d	e	f	g	h	t	
36..	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
36..	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
36..	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57
36..	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76
36..	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95
36..	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114
36..	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133
36..	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152
36..	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171
36..	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190
36..	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209
36..	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228
36..	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247
36..	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266
36..	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285
36..	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304
36..	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323
36..	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342
36..	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361
36..	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380
36..	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399
36..	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418
36..	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437
36..	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456
36..	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475
36..	476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494
36..	495	496	497	498	499	500	501	502	503	504	505	506	507	508	509	510	511	512	513
36..	514	515	516	517	518	519	520	521	522	523	524	525	526	527	528	529	530	531	532
36..	533	534	535	536	537	538	539	540	541	542	543	544	545	546	547	548	549	550	551
36..	552	553	554	555	556	557	558	559	560	561	562	563	564	565	566	567	568	569	570
36..	571	572	573	574	575	576	577	578	579	580	581	582	583	584	585	586	587	588	589
36..	590	591	592	593	594	595	596	597	598	599	600	601	602	603	604	605	606	607	608
36..	609	610	611	612	613	614	615	616	617	618	619	620	621	622	623	624	625	626	627
36..	628	629	630	631	632	633	634	635	636	637	638	639	640	641	642	643	644	645	646
36..	647	648	649	650	651	652	653	654	655	656	657	658	659	660	661	662	663	664	665
36..	666	667	668	669	670	671	672	673	674	675	676	677	678	679	680	681	682	683	684
36..	685	686	687	688	689	690	691	692	693	694	695	696	697	698	699	700	701	702	703
36..	704	705	706	707	708	709	710	711	712	713	714	715	716	717	718	719	720	721	722
36..	723	724	725	726	727	728	729	730	731	732	733	734	735	736	737	738	739	740	741
36..	742	743	744	745	746	747	748	749	750	751	752	753	754	755	756	757	758	759	760
36..	761	762	763	764	765	766	767	768	769	770	771	772	773	774	775	776	777	778	779
36..	780	781	782	783	784	785	786	787	788	789	790	791	792	793	794	795	796	797	798
36..	799	800	801	802	803	804	805	806	807	808	809	810	811	812	813	814	815	816	817
36..	818	819	820	821	822	823	824	825	826	827	828	829	830	831	832	833	834	835	836
36..	837	838	839	840	841	842	843	844	845	846	847	848	849	850	851	852	853	854	855
36..	856	857	858	859	860	861	862	863	864	865	866	867	868	869	870	871	872	873	874
36..	875	876	877	878	879	880	881	882	883	884	885	886	887	888	889	890	891	892	893
36..	894	895	896	897	898	899	900	901	902	903	904	905	906	907	908	909	910	911	912
36..	913	914	915	916	917	918	919	920	921	922	923	924	925	926	927	928	929	930	931
36..	932	933	934	935	936	937	938	939	940	941	942	943	944	945	946	947	948	949	950
36..	951	952	953	954	955	956	957	958	959	960	961	962	963	964	965	966	967	968	969
36..	970	971	972	973	974	975	976	977	978	979	980	981	982	983	984	985	986	987	988
36..	989	990	991	992	993	994	995	996	997	998	999	1000	1001	1002	1003	1004	1005	1006	1007
36..	1008	1009	1010	1011	1012	1013	1014	1015	1016	1017	1018	1019	1020	1021	1022	1023	1024	1025	1026
36..	1027	1028	1029	1030	1031	1032	1033	1034	1035	1036	1037	1038	1039	1040	1041	1042	1043	1044	1045
36..	1046	1047	1048	1049	1050	1051	1052	1053	1054	1055	1056	1057	1058	1059	1060	1061	1062	1063	1064
36..	1065	1066	1067	1068	1069	1070	1071	1072	1073	1074	1075	1076	1077	1078	1079	1080	1081	1082	1083
36..	1084	1085	1086	1087	1088	1089	1090	1091	1092	1093	1094	1095	1096	1097	1098	1099	1100	1101	1102
36..	1103	1104	1105	1106	1107	1108	1109	1110	1111	1112	1113	1114	1115	1116	1117	1118	1119	1120	1121
36..	1122	1123	1124	1125	1126	1127	1128	1129	1130	1131	1132	1133	1134	1135	1136	1137	1138	1139	1140
36..	1141	1142	1143	1144	1145	1146	1147	1148	1149	1150	1151	1152	1153	1154	1155	1156	1157	1158	1159
36..	1160	1161	1162	1163	1164	1165	1166	1167	1168	1169	1170	1171	1172	1173	1174	1175	1176	1177	1178
36..	1179	1180	1181	1182	1183	1184	1185	1186	1187	1188	1189	1190	1191	1192	1193	1194	1195	1196	1197
36..	1198	1199	1200	1201	1202	1203	1204	1205	1206										

Exemples de mauvaises numérisations (suite)

Google Books

(<http://books.google.com/books?id=XmpbAAAQAAJ>)

Livre numérisé à Oxford (Bodleian library) le 26 février 2009



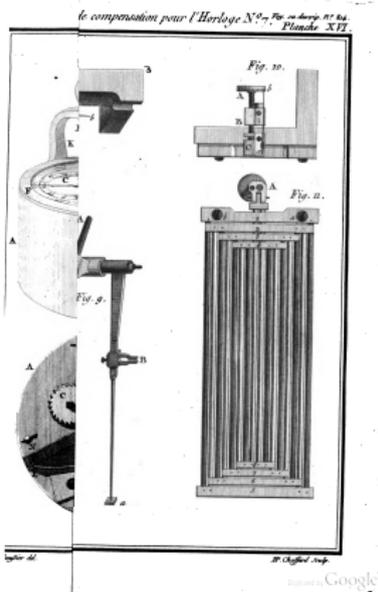
Digitized by Google

Exemples de mauvaises numérisations (suite)

Google Books

(<http://books.google.com/books?id=RVtDAAAACAAJ>)

Livre numérisé à l'université de Gand (Belgique) le 8 juillet 2010



Exemples de mauvaises numérisations (suite)

Ici, un exemple de numérisation par couches : certaines zones sont plus nettes que d'autres. C'est probablement une numérisation basée sur le format JPEG2000.

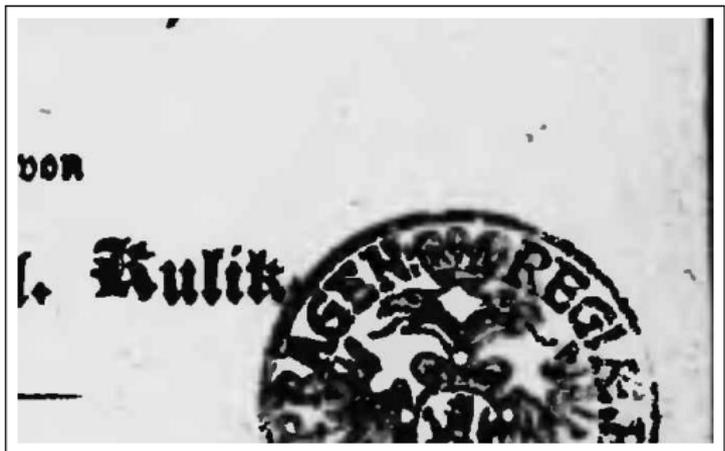
<http://kramerius.nkp.cz/kramerius/MShowMonograph.do?id=25654>



- vue de la première page de l'ouvrage
- l'ouvrage est présenté en images DJVU, avec possibilité de produire des PDFs partiels
- la numérisation par couches est ici un échec complet

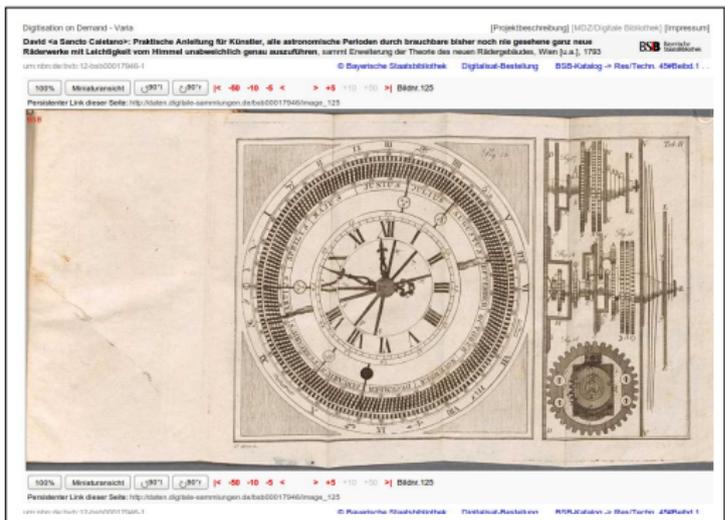
Exemples de mauvaises numérisations (suite)

Le JPEG2000 est souvent mal utilisé (notamment par Archive.org) :



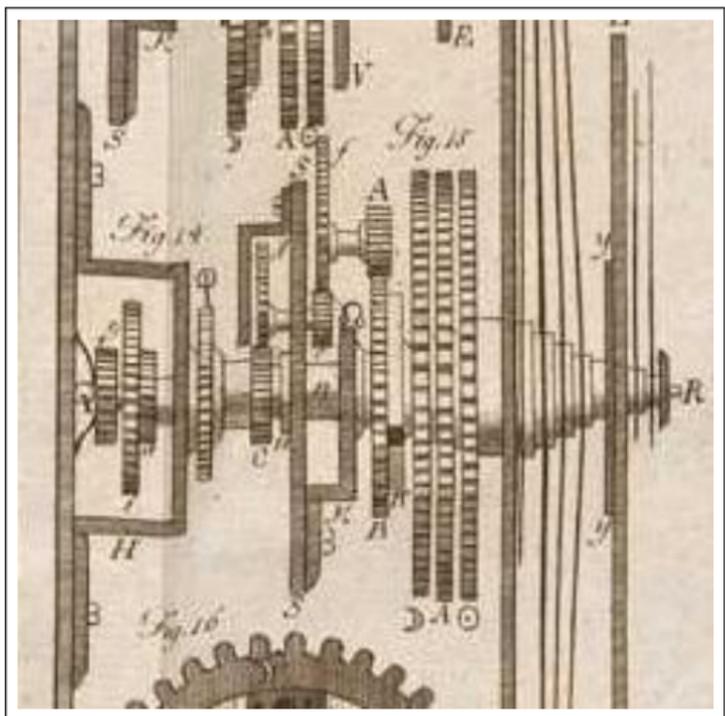
- détail de la vue précédente
- **il ne faut prendre du JPEG2000 que pour une bonne raison et ici, il y a mauvaise utilisation !**
- le flou sélectif n'est pas ce que les utilisateurs recherchent, sauf peut-être ceux qui ont des verres progressifs ;
- le format JPEG2000 est utilisé ici avec beaucoup de pertes.

Exemples de mauvaises numérisations (suite)



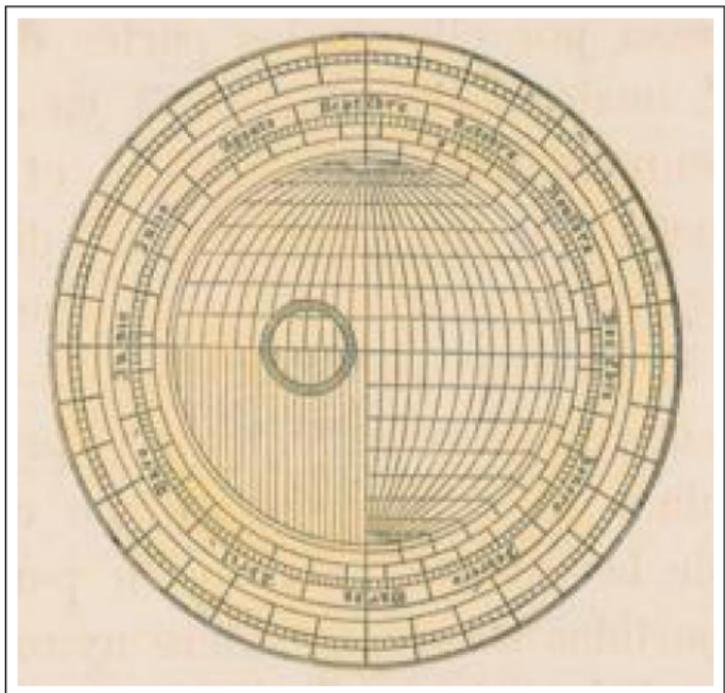
- une figure d'un ouvrage de Munich
- résolution maximale affichée

Exemples de mauvaises numérisations (suite)



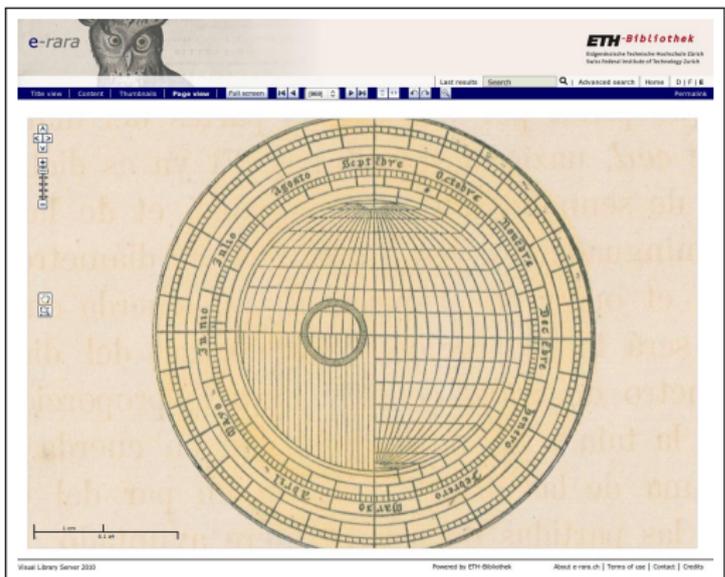
- détail de la vue précédente
- la résolution affichée est clairement trop faible

Exemples de mauvaises numérisations (suite)



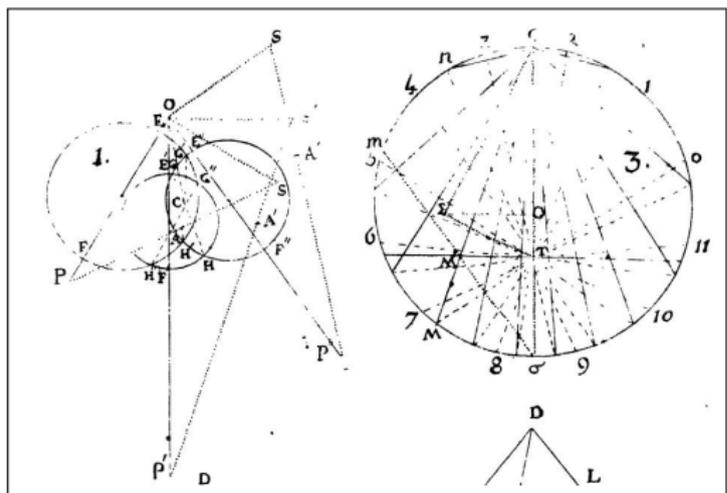
- projet e-rara (Zurich)
- une figure du livre *Libros del Saber*, volume 3, p. 143
- inscriptions totalement illisibles dans le PDF téléchargé

Exemples de mauvaises numérisations (suite)



- en ligne, par contre, on peut obtenir une meilleure qualité

Exemples de mauvaises numérisations (suite)

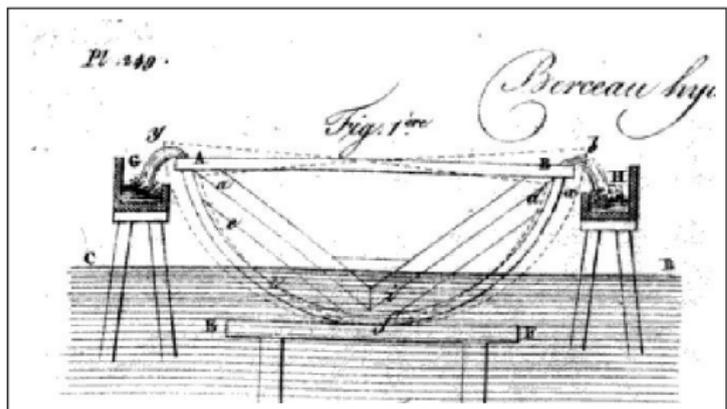


- projet NUMDAM (<http://www.numdam.org>, numérisation de revues mathématiques)
- une figure géométrique (Annales de Gergonne, volume 17)
- une grande partie des traits ont disparu du fait d'une résolution insuffisante

Exemples de mauvaises numérisations (suite)

Conservatoire numérique des Arts et Métiers
(CNUM, <http://cnum.cnam.fr>)

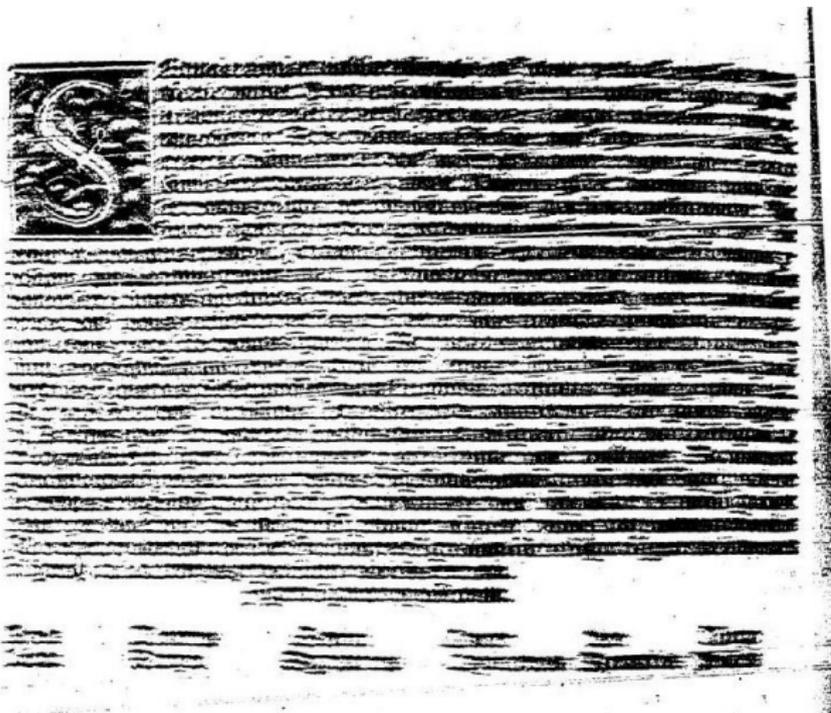
Ici, extrait du *Bulletin de la Société d'Encouragement*, 1823.



- numérisation de figure avec résolution insuffisante
- le numéro de la planche (249) est impossible à lire avec certitude
- planche non alignée
- une nouvelle numérisation est indispensable

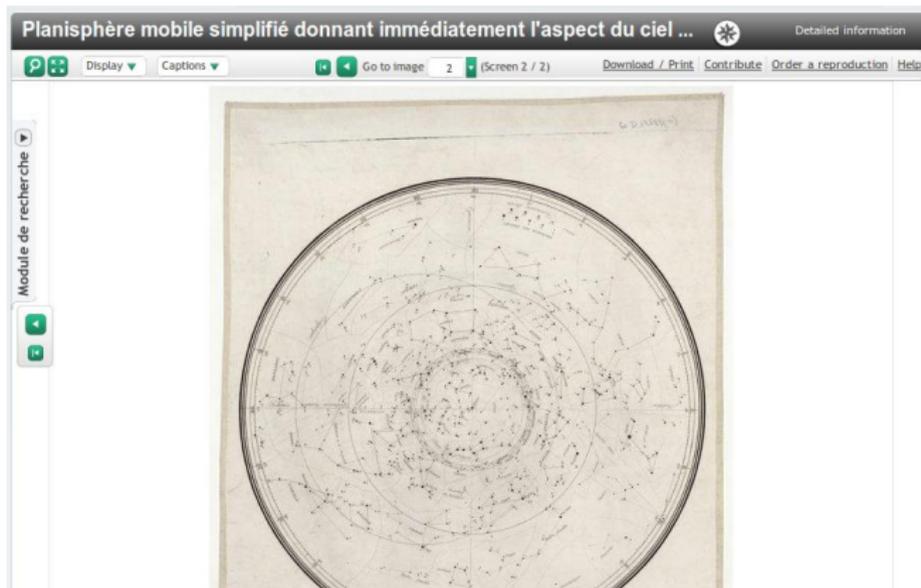
Exemples de mauvaises numérisations (suite)

Gallica : *Astrolabium planum in tabulis ascendens*, 1494



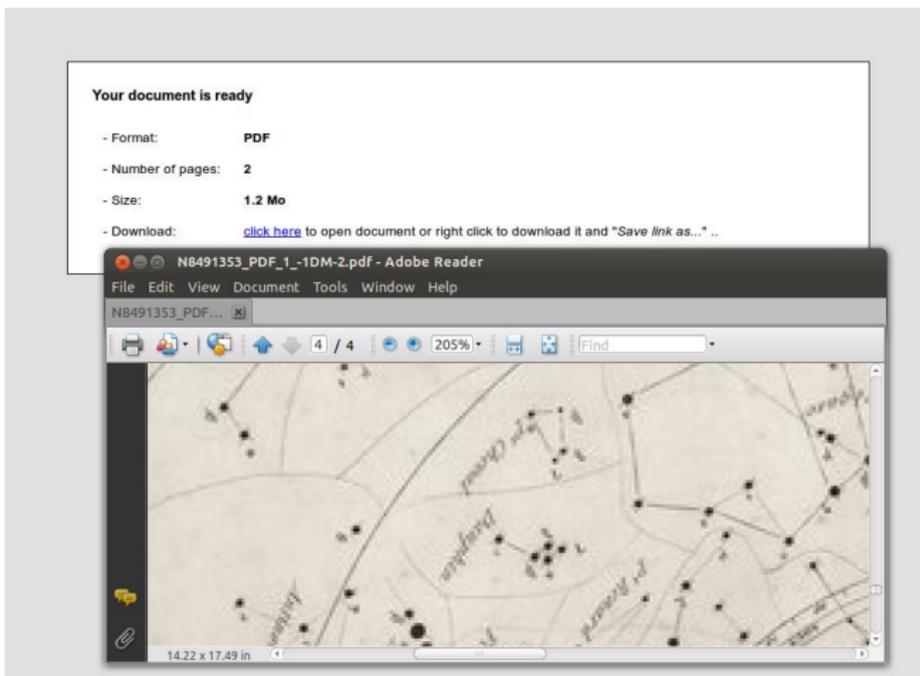
Exemples de mauvaises numérisations (suite)

Gallica : *Planisphère mobile simplifié*, 1839



Exemples de mauvaises numérisations (suite)

Gallica : *Planisphère mobile simplifié*, 1839



Exemples de mauvaises numérisations (suite)

Archive.org : 數理精蘊 (Shuli Jingyun, 1723)

<http://www.archive.org/details/06076323.cn>

五〇三〇一	五〇二八一	五〇二六一	五〇二四一	五〇二二一	五〇二〇一	五〇一八一	五〇一六一	欽 定 四 庫 全 書
五〇三〇二	五〇二八二	五〇二六二	五〇二四二	五〇二二二	五〇二〇二	五〇一八二	五〇一六二	
五〇三〇三	五〇二八三	五〇二六三	五〇二四三	五〇二二三	五〇二〇三	五〇一八三	五〇一六三	
五〇三〇四	五〇二八四	五〇二六四	五〇二四四	五〇二二四	五〇二〇四	五〇一八四	五〇一六四	
五〇三〇五	五〇二八五	五〇二六五	五〇二四五	五〇二二五	五〇二〇五	五〇一八五	五〇一六五	
五〇三〇六	五〇二八六	五〇二六六	五〇二四六	五〇二二六	五〇二〇六	五〇一八六	五〇一六六	
五〇三〇七	五〇二八七	五〇二六七	五〇二四七	五〇二二七	五〇二〇七	五〇一八七	五〇一六七	
五〇三〇八	五〇二八八	五〇二六八	五〇二四八	五〇二二八	五〇二〇八	五〇一八八	五〇一六八	

(probablement d'après microfilm)

Exemples de mauvaises numérisations (suite)

Archive.org : 數理精蘊 (Shuli Jingyun, 1723)

La même chose en mieux

五〇三〇一	五〇二八一	五〇二六一	五〇二四一	五〇二二一	五〇二〇一	五〇一八一	五〇一六一
二四三 二〇七	六五三 七七		一六七四七 三		二九五三 一七	三八九 一八九	四八七 一〇三
五〇三〇二	五〇二八二	五〇二六二	五〇二四二	五〇二二二	五〇二〇二	五〇一八二	五〇一六二
三五九三 一四	八一 六二	八三七七 六	二五一二 二	二五一 二	二七八九 一八	二二八 二二	三五八三 一四
五〇三〇三	五〇二八三	五〇二六三	五〇二四三	五〇二二三	五〇二〇三	五〇一八三	五〇一六三
二六九 一八七	三三三 一五一		一〇六九 四七	一六七四 三	八二三 六一	四六九 一〇七	七二七 六九
五〇三〇四	五〇二八四	五〇二六四	五〇二四四	五〇二二四	五〇二〇四	五〇一八四	五〇一六四
二六二 一九二	九六七 五二	二四四 二〇六	二三七 二二二	二九二 一七二	三〇八 一六三	二四六 二〇四	一二五四 四
五〇三〇五	五〇二八五	五〇二六五	五〇二四五	五〇二二五	五〇二〇五	五〇一八五	五〇一六五
一〇〇六 一五	四四五 一三三	一一一七 四五	七七三 六五	二四五 二〇五	三三四 一五	一〇〇三 七五	三九五 二二七
五〇三〇六	五〇二八六	五〇二六六	五〇二四六	五〇二二六	五〇二〇六	五〇一八六	五〇一六六
二五一五 三	二八九 一七四	六一三 八二	二五九 一九四	七六一 六六	一九三一 二六	一〇九 四六	九二九 五四
五〇三〇七	五〇二八七	五〇二六七	五〇二四七	五〇二二七	五〇二〇七	五〇一八七	五〇一六七
四〇九 一三		三〇一 一六七	一八六一 二七			一六七二 九三	三二七 三二

Ces numérisations suffisent-elles ?

- si le but est d'avoir un aperçu, ces numérisations suffisent ;
- avec la dernière numérisation, on voit que c'est du chinois, la majorité des gens s'en contenteront ;
- mais pour vraiment lire ce qui est écrit, cela ne suffit pas ;
- de plus, il est facile de concevoir que **le problème n'est pas un problème d'affichage, mais un problème au niveau de la source** ;
- **Avec certaines numérisations, la lecture devient de l'archéologie !**
- lire un document numérique doit-il être pénible ?

Niveau de qualité et commanditaires

La qualité moyenne est clairement liée à l'organisme responsable de la numérisation.

- Google books et Internet Archive :
 - qualité moyenne et quelquefois très faible
 - jamais très bonne
- bibliothèques allemandes
 - qualité moyenne très bonne
 - quelquefois moins sur des aspects précis
- certains organismes produisent des numérisations et interfaces de très faible qualité ou douteuses :
 - Digital Library of India
 - Universal Digital Library (Carnegie Mellon, numérisations sous-traitées en Inde et en Chine)
- pour beaucoup de sites, **la qualité augmente avec le temps**

Le cas de Google books (1)

- objectif uniquement économique ;
- les ouvrages numérisés doivent être accessibles dans la plate-forme *e-books* de Google ;
- la préoccupation de Google n'est pas le patrimoine, mais le **contrôle de l'information écrite** ;
- la qualité de numérisation est faible (mais s'améliore, par exemple par le **passage à la couleur mi-2009**) ;
- un nombre important de bibliothèques sont tombées dans le « piège » Google books ;
- limites d'accès même pour ouvrages libres de droits, variant suivant les pays ;

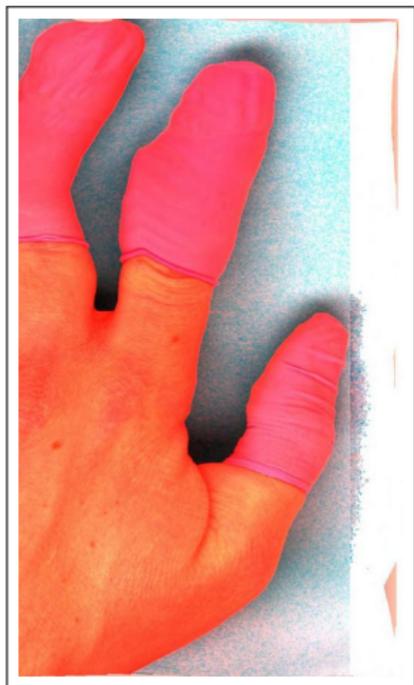
Le cas de Google books (2)

Les ouvrages numérisés par Google ont de nombreux défauts :

- des pages peuvent être floues, des doigts peuvent être visibles, etc.
- les planches pliées des ouvrages anciens ne sont jamais dépliées ;
- certains de ces défauts ne sont pas accidentels, mais correspondent à des consignes ;
- ceux qui numérisent pour Google ont pour consigne de ne pas déplier les planches ; si ce n'était pas le cas, on trouverait quelquefois des planches dépliées ;
- Google ne se gêne pas pour vendre des ouvrages incomplètement scannés.

Le cas de Google books (3)

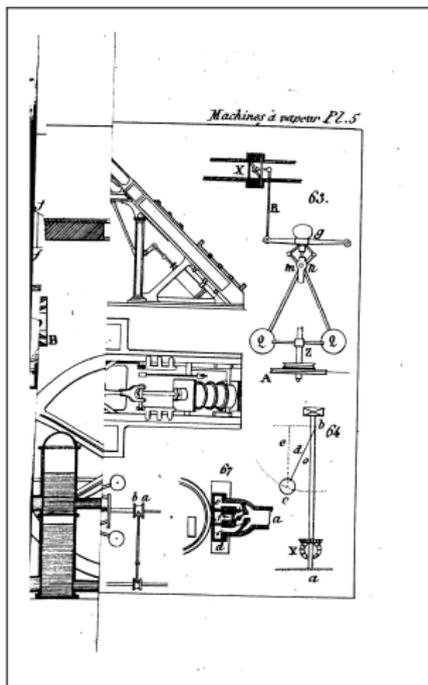
Les doigts d'un opérateur



Ozanam : *Récréations mathématiques*, 1696,
Google à la bibliothèque de Munich

Le cas de Google books (5)

Planches non dépliées

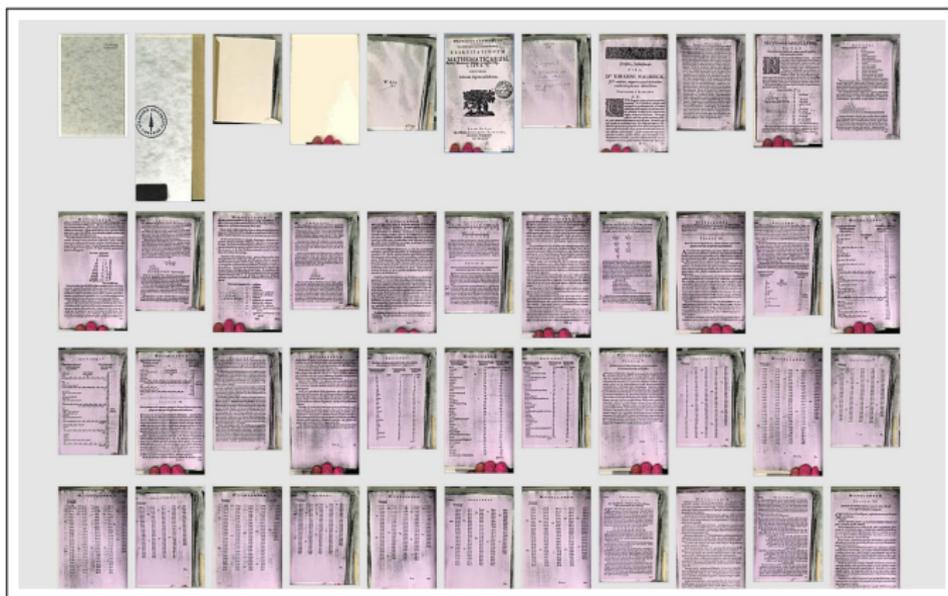


*Traité complet des machines à vapeur
appliquées à l'industrie, 1838*

Google à bibliothèque municipale de Bavière,
2012

Le cas de Google books (6)

Encore des doigts



Schooten, *Exercitationum mathematicarum*, 1657,
Google à Stanford

Le cas de Google books (4')

Encore plus de doigts



Le cas de Google books (7)

Il y a une méconnaissance assez générale des problèmes techniques de Google, même dans les bibliothèques :

- dans leurs livres sur Google, ni Jean-Noël Jeanneney (ancien directeur de la BNF, critique à l'égard de Google), ni Bruno Racine (actuel directeur de la BNF, plutôt ouvert au dialogue avec Google), ne mentionnent les problèmes de qualité de Google, et probablement ne les connaissent pas ;
- de manière générale, indépendamment des erreurs de numérisation, l'insuffisance de la résolution n'est jamais évoquée ;

Le cas de Google books (8)

- Alain Jacquesson (ancien directeur de la bibliothèque de Genève), dans son livre « Google Livres et le futur des bibliothèques numériques » (2010), consacre deux pages au contrôle de qualité ; pour l'auteur, les erreurs de Google sont toutes accidentelles et il croit à l'offre de Google de les corriger ; il semble croire que les ouvrages avec planches dépliantes n'entrent pas dans la chaîne de numérisation ; il semble aussi ignorer que les ouvrages mal numérisés sont mis en vente ;
- Jacquesson parle des « erreurs de manipulation lors de la prise de vue » et ajoute « Google donne (...) des directives précises à ses opérateurs de numérisation ».

Le cas de Google books (9)

Google a introduit un certain nombre d'**innovations techniques**, mais ce ne sont pas des miracles :

- Google a déposé un brevet pour la correction des défauts de courbure engendrés par la non-planéité des ouvrages épais ;
- on constate sur certaines vues des corrections visiblement informatiques ; ces corrections entraînent de telles déformations que **le résultat est presque pire que sans corrections** ;
- le moins qu'on puisse dire, c'est que le brevet de Google n'est pas au point ou pas bien appliqué !

Il semble que les bibliothèques et autres professionnels du livre soient très peu au fait des problèmes techniques de Google.

Le cas de Google books (10)

Les planches dépliantes :

- selon Alain Jacquesson, les ouvrages qui en contiennent sont exclus de la numérisation (p. 60 : « Les ouvrages comportant des figures à déplier sont aussi rejetés ») ;
- le fait que de tels ouvrages sont numérisés est alors ou bien une erreur de Google qui les prend tout de même, ou bien une erreur de la bibliothèque fournisseuse qui fournit les ouvrages ;
- par cohérence, Google devrait détruire une numérisation dès qu'elle contient des planches qui conduiraient à une numérisation partielle ;

Google préfère donc numériser mal (et sciemment), que pas du tout.

Le problème des pseudo-éditeurs (recycleurs de livres) (1)

- accords probables entre Google et de « faux » éditeurs comme bibliobazaar = BiblioLife = Nabu Press, Kessinger, Pranava, Woods Press, etc. ;
- les numérisations sont celles de Google, mais sans le logo Google ; ce sont donc des réutilisations autorisées ;
- l'impression n'est faite qu'à la demande, mais les catalogues sont remplis de ces livres inexistantes ;
- Google se cache peut-être derrière les noms précédents ;
- les versions recyclées font *ombrage* aux versions d'origine ;
- mise en avant de ces versions (en contradiction avec la volonté affichée de Google de combattre les « fermes de contenus »), mais en accord avec la collusion probable ;

Le problème des pseudo-éditeurs (2)

- ces « faux » éditeurs ne semblent pas se recouper et semblent se partager le marché ; un livre donné n'est pas publié par plus d'un de ces éditeurs ;
- problème de pérennité : lorsqu'une version ancienne est réimprimée par un de ces pseudo-éditeurs, la version d'origine peut ne plus être accessible dans Google books (pour un exemple, voir la page Wikipedia de Kessinger Publishing) ;

Numéro de notice : [077743806](#)

Titre : [Aesthetic measure](#) [Texte Imprimé] / George D. Birkhoff

Alphabet du titre : latin

Auteur(s) : [Birkhoff, George David \(1884-1944\)](#). Auteur

Date(s) : [s.d.]

Langue(s) : anglais

Pays : Etats-Unis

Editeur(s) : [S.l.] : [Kessinger](#), [s.d.]

Description : xii, [1] p., 2 l., 3-225, [1] p. : illus. (incl. music) XXIII pl. (part col.) diagrs. ; 30 cm

Notes : Reprod. en fac-similé de l'ed. Harvard University Press de 1933

Sujets : [Proportions \(art\)](#)
[Musique](#) -- [Philosophie](#)
 BH201. .B5

Origine de la notice : AKR

- une notice SUDOC d'un ouvrage recyclé par Kessinger
- petit à petit, les vrais livres vont être remplacés par des copies recyclées de mauvaise qualité (invasion des profanateurs ?)

Le problème des pseudo-éditeurs (3)

Exemple de recherche sur AddALL avec 52 réponses, dont 22 sont des impressions à la demande de Woods Press :

Save the Info	Sort Asc TITLE Sort Desc	Asc AUTHOR Desc	Asc PRICE USD Desc	Asc SITE Desc	Asc DEALER Desc	DESCRIPTION
	1 Traite D'Horlogerie Moderne Theorique Et Pratique Buy it!	Claudius Saunier	24.41	AmazonCA	AmazonCA	NEW, Woods Press 2010-11 Paperback 144650672X 24.65 CAD to USD is calculated base on 1 CAD = 0.99018 USD
	2 Traite D'Horlogerie Moderne Theorique Et Pratique Buy it!	Claudius Saunier	39.80	AmazonUK	AmazonUK	NEW, Unknown 2010-11 Paperback 144650672X 25.46 GBP to USD is calculated base on 1 GBP = 1.56336 USD
	3 TraitÃ© d'horlogerie moderne thÃ©orique et pratique (French Edition) Buy it!	Claudius Saunier	40.95	Amazon	Amazon	NEW, Woods Press 2010-11-01 Paperback 144650672X
	4 Traite D'Horlogerie Moderne Theorique Et Pratique Buy it!	Claudius Saunier	41.97	AmazonFR	AmazonFR	NEW, Woods Press 2010-11 BrochÃ© 144650672X 31.31 EUR to USD is calculated base on 1 EUR = 1.34055 USD
	5 Traite D'Horlogerie Moderne Theorique Et Pratique Buy it!	Saunier, Claudius	45.33	Abebooks	Paperbackshop-US	[publisher: WOODS PR 01/01/2012] New print on demand book. Shipped from US. This item is printed on demand. [Aurora, IL, U.S.A.]
	6 Traite D'Horlogerie Moderne Theorique Et	Claudius Saunier	45.79	AmazonFR	AmazonFR	USED, Woods Press 2010-11 BrochÃ© 144650672X

Mauvaises numérisations papier

Le problème des mauvaises numérisations ne concerne pas uniquement les ouvrages en ligne :

- certains ouvrages nouveaux sont de qualité insuffisante, même chez des éditeurs prestigieux ;
- certaines réimpressions montrent une absence de maîtrise de la numérisation.

Le cas de Cambridge University Press

En 2010, Cambridge University Press a réimprimé l'ouvrage de Henderson, *Life of James Ferguson*, 1867.

Il y a ici deux problèmes :

- on ne comprend pas très bien pourquoi CUP a réimprimé l'édition de 1867, alors qu'il y a eu une édition corrigée en 1870 ;
- **un grand nombre de figures apparaissent floues**, comme si leur traitement avait mis en œuvre une résolution insuffisante ; curieusement, ce n'est pas le cas de toutes les figures.

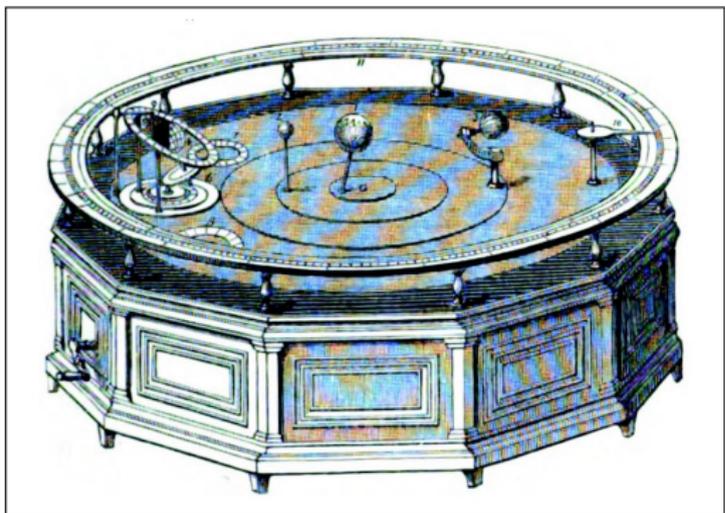
Le cas de Cambridge University Press (2)

On peut comparer les figures de 1870 et de la réimpression de 2010 sur Google Books, et aussi comparer les figures de 1870 en ligne et en PDF, qui ne sont pas exactement les mêmes.



- figure page 73
- version de 1870 en ligne

Le cas de Cambridge University Press (3)



- version de 1870 en PDF
- plus contrastée que la version en ligne

Le cas de Cambridge University Press (4)

preserves its parallelism during its annual course ; thus exhibit-



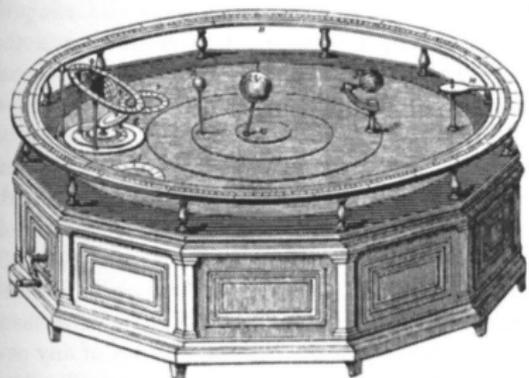
Ferguson's Orrery.

ing her different length of days and nights; her change of sea-

- version de 1867 (numérisée à partir d'un original)
- même netteté que la version de 1870

Le cas de Cambridge University Press (5)

preserves its parallelism during its annual course ; thus exhibit-



Ferguson's Orrery.

ing her different length of days and nights; her change of sea-

- version de 2010 (numérisée à partir d'un original)
- nettement plus floue que la précédente : la figure est tramée
- ce problème est aussi visible sur la réimpression telle qu'elle apparaît sur Google books

Le rapport Tessier

Rapport sur la numérisation du patrimoine écrit
(disponible sur www.culture.gouv.fr)

- rapport commandé suite au rapprochement entre la BNF et Google
- rapport remis le 12 janvier 2010 par Marc Tessier
- fait l'état des lieux de la numérisation
- préconise des solutions

À propos de la « médiocrité de la numérisation opérée par Google », le rapport signale (p. 51) que la BNF est maintenant passée de 300 à 400dpi. C'est mieux, mais ce n'est pas encore durable !

Problèmes de résolution

Il y a essentiellement deux types de problèmes de résolution :

- les problèmes que l'on rencontre le plus fréquemment sont liés à une résolution trop faible
- mais on a aussi le contraire, qui est moins connu.

Problèmes de surnumérisation (1)

Exemple d'une planche d'un ouvrage numérisé à Strasbourg (cote H349), numérisé à environ 400 ou 500dpi.



Problèmes de surnumérisation (2)

Détail de la précédente vue :



Problèmes de surnumérisation (3)

Détail de la précédente vue :



- à gauche des caractères : un peu **vert** ; à droite : un peu **jaune** ;
- phénomène d'aberration chromatique.
- le vert est toujours du même côté des caractères, ce qui exclut un artéfact de compression JPEG

Problèmes de surnumérisation (4)

- comme l'exemple précédent l'a montré, certaines bibliothèques numérisent à une résolution trop grande (au-delà des possibilités techniques)
- les numérisations introduisent alors des défauts, par exemple dus à l'optique (aberrations chromatiques, etc.)
- le matériel impose une limite
- le numériseur doit donc avoir quelques notions physiques, ce qui est rarement le cas

surnumérisation = fausse qualité

Évaluation de la qualité de la numérisation

Comment évaluer la qualité de la numérisation ?

Idée : **demander aux utilisateurs.**

- il y a différents types d'utilisateurs, qui se scindent majoritairement en deux groupes :
 - ceux qui feuillentent, et
 - ceux qui utilisent vraiment un ouvrage
- or les bibliothèques numériques sont vues comme des extensions du web et le mode de consultation est le même : c'est le règne de la superficialité ;
- les utilisateurs qui feuillentent sont donc majoritaires et sont en général satisfaits ; faut-il pour autant se baser sur eux pour les choix de qualité ?

Évaluation de la qualité de la numérisation

- la satisfaction d'un utilisateur doit être modulée par son besoin ;
- le besoin peut être connu ou non ;
- certains besoins n'apparaissent qu'à l'utilisation, notamment parce qu'ils sont vus comme acquis ;
- un besoin de rapidité ne devient par exemple explicite que lorsque le service fourni est lent ; **ce n'est pas parce que l'utilisateur n'exprime pas un certain besoin qu'il ne l'a pas ;**
- connaître les besoins des utilisateurs et leurs évolutions n'est donc pas toujours simple.

Évaluation de la qualité de la numérisation

L'interrogation de l'utilisateur présente donc un risque :

- il est tentant d'évaluer les besoins comme étant ceux de la majorité ;
- les utilisateurs qui sont critiques sont *de facto* minoritaires, puisqu'il y a une majorité d'utilisateurs qui feuillentent (combien d'entre nous contactent les sites lorsque la mise en page ou le contenu laissent à désirer ?) ;
- les utilisateurs critiques sont alors marginalisés : on leur dit que leur opinion n'est pas représentative.

Évaluation de la qualité de la numérisation

On peut aussi évaluer la qualité de numérisation par rapport au **confort de lecture**. Le confort visuel dépend assez peu du support : parchemin, papyrus, tablette d'argile, papier, etc.

Nous regardons toujours avec les mêmes yeux !

Dans le cas de la numérisation :

- le confort est plus ou moins proche du confort de l'original
- si on devine certains caractères, ce n'est pas suffisant ;
- chaque caractère doit pouvoir être identifié sans ambiguïté ;
- le document doit permettre l'extraction de toutes les informations ;
- les images et schémas doivent montrer tous les détails et être suffisants pour une republication

Qualité de stockage et qualité fournie (1)

- il peut être nécessaire de distinguer la qualité de stockage de la qualité fournie à l'utilisateur ;
- on peut par exemple numériser à 600dpi, et ne mettre en ligne au mieux que des images à 300dpi ; (c'est peut-être ce que fait Google, mais je doute que les éventuelles versions plus fines soient dépourvues des erreurs visibles en ligne)
- par la suite, les accès devenant plus rapides, on pourra éventuellement basculer vers une plus grande qualité ;

Qualité de stockage et qualité fournie (2)

- il est souhaitable qu'il y ait une certaine transparence ;
- l'utilisateur doit être informé de la qualité de stockage, même s'il n'y a qu'occasionnellement accès ;

Qualité de stockage et qualité fournie (3)

- Lorsqu'une vue en ligne est de qualité trop faible, on ne doit pas dire « c'est normal, la version en ligne est dégradée ». Car alors se poserait la question : « pourquoi la mettre en ligne ? » Pour que les utilisateurs payent pour un meilleur accès ?
- La vraie question est : pourquoi ceux qui ont mis en ligne n'ont-ils pas compris 1) qu'il fallait une meilleure version en ligne, et 2) que c'était possible ?

La réponse est que ceux qui numérisent et mettent en ligne ne comprennent pas les besoins des utilisateurs. De plus, mettre en ligne une version de meilleure qualité n'est pas nécessairement plus lourd pour l'utilisateur, lorsque le parcours n'est qu'en ligne. La qualité ne devrait avoir un coût qu'au téléchargement.

Qualité de stockage et qualité fournie (4)

Les différents niveaux de qualité imposent donc assez naturellement une structuration des documents :

- on peut rendre accessible la totalité du document avec une bonne qualité, mais à un coût (volume, temps de téléchargement) élevé ;
- on peut aussi (en plus) segmenter le document pour en permettre des téléchargements partiels (prédéfinis ou à la demande) ;
- enfin, on peut prévoir des qualités variables suivant les parties du document ; des planches peuvent par exemple être numérisées avec une meilleure qualité que le texte.

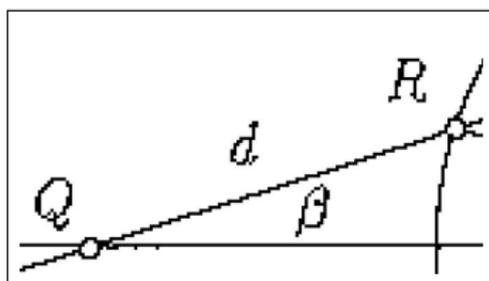
Dans tous les cas, les bibliothèques ne peuvent pas se contenter de mettre en ligne des versions dégradées sans avoir évalué les besoins des utilisateurs.

Qualité de stockage et qualité fournie (5)

On constate que certains sites accroissent la qualité avec le temps, dans certains cas à partir des mêmes numérisations internes. Un exemple intéressant est JSTOR.

Voici des extraits de l'article d'Arthur Baragar paru dans la revue *American Mathematical Monthly* en 2002. Le PDF a été généré en 2007, au moment du téléchargement du fichier (<http://www.vcharkarn.com/uploads/2/2557.pdf>).

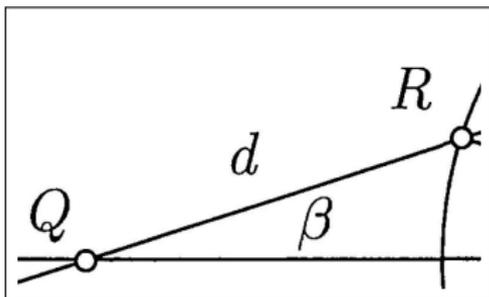
1. INTRODUCING
 mathematicians have
 provocative. To
 applications to



Qualité de stockage et qualité fournie (6)

En 2011, JSTOR produit des fichiers de meilleure qualité :

1. INTROD
maticians ha
provocative.
applications



Qualité de stockage et qualité fournie (7)

L'examen des informations des fichiers PDF montre qu'ils ont été produits différemment, peut-être à partir de la même source :

en 2007

Created: 02/05/2007 05:08:04 PM
Modified: 02/05/2007 05:08:04 PM
Application:
Advanced
PDF Producer: iText 1.3 by lowagie.com (based on itext-paulo-153)
PDF Version: 1.4 (Acrobat 5.x)

(produit au téléchargement)

en 2011

Created: 03/06/2008 09:01:08 AM
Modified: 03/07/2008 11:54:20 AM
Application: PDFplus
Advanced
PDF Producer: Atypon Systems, Inc.
PDF Version: 1.4 (Acrobat 5.x)

(produit trois ans avant)

En outre, on note que les deux PDF ont des tailles comparables.

Résolution nécessaire

Comment déterminer la résolution nécessaire ?

- le numériseur doit comprendre le lien entre différentes entités, et avant tout entre le lecteur et le livre ;
- les caractères ont une certaine finesse ;
- l'utilisateur regarde avec ses yeux ;
- un texte imprimé à 300dpi présente des irrégularités visibles à l'œil nu, à 10 ou 20 cm ;
- dans les mêmes conditions, un texte imprimé à 600dpi apparaît lisse ;
- ceci est compatible avec le pouvoir séparateur de l'œil qui est d'environ 1' ; à 50 cm, cela correspond à environ 300dpi ;
- ceci justifie une **résolution durable de 600dpi pour du texte** ;

Résolution nécessaire et contenu informationnel

La résolution nécessaire dépend surtout du contenu informationnel du document :

- un texte est destiné à être lu à 50 cm, 600dpi doit donc suffire ;
- si le texte doit être agrandi et si d'autres détails sont importants, une résolution plus importante peut être nécessaire ;

Résolution nécessaire et contenu informationnel

- une photographie d'un ouvrage peut en général être numérisée à 600dpi, même si elle est tramée ; toutefois, dans ce cas, une résolution inférieure sera peu gênante, car l'utilisateur ne verra pas l'impact de la faible qualité de numérisation sur un original déjà de faible qualité ; la question se pose de savoir si l'on veut conserver la qualité, même mauvaise, des vues initiales ;
- certaines photographies contiennent bien plus de détails que n'en fournissent 600dpi ; certains tirages contacts anciens peuvent nécessiter 1000dpi ou plus, et les négatifs peuvent requérir 3000dpi ;
- la résolution nécessaire dépend aussi de l'utilisation : une photographie n'est pas forcément destinée à être vue en grandeur réelle à 50 cm des yeux ;
- mais pour des imprimés, 600dpi devrait presque toujours convenir.

Y a-t-il une limite des besoins ?

Un ancien responsable de la numérisation me rappelait récemment l'évolution des formats de stockage :

- 1 argentique N&B puis couleur
- 2 photocopie
- 3 microfiche
- 4 microfilm N&B puis couleur
- 5 vidéodisque
- 6 TIFF basse résolution
- 7 etc.

Beaucoup de responsables semblent percevoir l'évolution comme un éternel remplacement de technologies, avec des longévités et des capacités croissantes.

Y a-t-il une limite des besoins ?

Ce que la liste précédente suggère, c'est que les besoins croissent.

- Les bibliothèques et responsables de numérisation se sont donc faits à l'idée de la non-durabilité.
- Pour certains responsables de numérisation, le concept de durabilité n'a ainsi pas un sens très précis.

Mais, comme me le rappelait un médecin récemment : si les capacités des ordinateurs croissent, il n'en est pas ainsi (ou très peu) avec nos cerveaux.

Et nos yeux restent aussi les mêmes.

Ce sont les yeux et l'écriture qui fixent les besoins limites.

Il y a une limite des besoins ! Limite = équivalence

- (cas général) si le but est de lire un texte :
Lorsque la numérisation d'un texte atteint une résolution telle que ce texte, présenté dans des conditions normales de lecture, ne fait plus apparaître d'aspérités, alors il n'est plus nécessaire d'aller au-delà pour le confort de lecture.
Cette résolution limite dépend uniquement du pouvoir de résolution de l'œil.
- si le but est d'extraire une autre information utile :
Il faut définir quelle est cette information utile et en déduire la résolution (ou un autre paramètre) nécessaire.

Ce sont ces considérations qui permettent de définir les critères minimums de numérisation durable.

A-t-on besoin du confort de lecture ?

On peut faire le parallèle avec les technologies des écrans :

- nous évoluons vers la télévision HD, des écrans d'ordinateur plus grands, des écrans de portables plus fins, des films en blu-ray, etc.
- ces technologies vont de plus en plus isoler les numérisations de faible qualité ;
- les lecteurs d'aujourd'hui et de demain voudront de plus en plus la qualité d'image ;

Il ne fait donc pas de doute que si le confort n'est pas pris en compte, la durabilité ne pourra pas suivre.

Retour sur la qualité de stockage et la qualité fournie (1)

- nous savons maintenant qu'il y a une résolution limite pour les usages courants, c'est-à-dire les livres lus avec les yeux (et pas une loupe) ; il n'est donc pas nécessaire d'aller au-delà de cette résolution (pour cet usage) ;
- ceci définit donc la qualité durable pour le stockage ; en-deça de cette qualité, la numérisation devra un jour être refaite ;
- la qualité fournie à l'utilisateur dépend du confort que l'on peut (ou est prêt à) fournir ;
- certains contenus nécessitent que la qualité fournie soit plus proche de la qualité de stockage que d'autres, c'est notamment le cas pour des exemples vus précédemment.

Retour sur la qualité de stockage et la qualité fournie (2)

L'utilisateur ne connaît souvent pas la qualité de stockage, mais si ce que fournissent les bibliothèques est de qualité *inutilisable* (comme dans les exemples vus plus haut), on peut se demander

- si les bibliothèques comprennent ce qu'elles font et comprennent les besoins des utilisateurs ;
- si un éventuel stockage interne est vraiment de qualité suffisante.

Pour éviter ces doutes, les bibliothèques doivent aux utilisateurs une transparence totale sur les moyens et les choix, tant techniques que politiques.

Couleur ou noir et blanc ?

Lorsqu'un document n'est imprimé qu'en noir sur du papier, on peut se demander si la numérisation en noir et blanc n'est pas suffisante.

- c'est vrai, mais le papier étant rarement blanc, l'apparence de la numérisation ne correspond pas à celle du document réel ; il y a donc un impact sur la sensation d'identité ;
- si seulement certaines pages sont numérisées en couleur, il y a un problème de cohérence ; ces numérisations ne seront pas viables à long terme ;
- Google met en ligne les numérisations en couleur depuis 2009 (mais le téléchargement reste en noir et blanc).

La plupart des grandes bibliothèques allemandes font aussi les numérisations en couleur, y compris au niveau des téléchargements.

Quantité ou qualité ?

Comme souvent, la qualité s'oppose à la quantité :

- ceux qui numérisent peu ont tendance à numériser mieux ;
- ceux qui numérisent beaucoup ont tendance à numériser moins bien ;

Se pose alors la question : est-il vraiment plus intéressant pour l'utilisateur d'avoir beaucoup de livres moyennement numérisés que peu de livres bien numérisés ?

Je pense que l'on essaie de nous faire croire que nous avons besoin de la quantité.

Explosion de l'information

- En quelques années, une quantité immense de documents sont devenus accessibles.
- Nous sommes passés en 10 ans de pratiquement 0 livres à 10 millions ou plus.
- Mais les besoins ne sont pas encore là.
- Et nous n'avons pas le temps d'exploiter ces ressources !
- Qui a besoin de 10 millions de livres ?
- Et de surcroit, non durables ?

On nous vend la quantité (et la mauvaise qualité), alors que nous pourrions nous contenter de beaucoup moins.

La quantité est un besoin économique (1)

Nous avons pu entendre des affirmations comme :

Google peut numériser en 5 ans ce qui nous aurait demandé 50 ans

Mais est-il **nécessaire** d'aller aussi vite ?

Nous avons (presque) tous un téléphone portable :

mais en avons-nous **vraiment** besoin ?

La quantité est un besoin économique (2)

- aller vite (et mal) est une nécessité économique
- ce n'est pas un besoin de l'utilisateur
- mais si un acteur va vite (et mal) et l'autre lentement (et bien), l'utilisateur ne perçoit pas l'avantage qualitatif ;
- Google force donc les autres projets à aller vite et à avoir une attitude commerciale alors qu'elle ne se justifie pas au niveau scientifique

Qualité et durabilité

- Concentrons-nous donc sur la qualité et la durabilité.
- Et précisons les critères de numérisation.

Critères minimums pour une numérisation durable

Nos recommandations sont les suivantes :

- au moins 600 dpi
- couleur
- pas de retouche des images
- le TIFF n'est pas nécessaire
- compression avec pertes non visibles
- pas d'aberrations (optiques ou autres) visibles
- planches dépliées
- originaux alignés
- certains supports photo (notamment d'anciens tirages contacts) nécessitent plus de 600 dpi

La qualité des interfaces

- nous allons maintenant faire un petit panorama des interfaces ;
- et relever les points positifs et négatifs.

Examen de quelques interfaces

Le problème de la qualité

- petit panorama critique
- Gallica, Strasbourg, Berlin, Goettingen, Dresden, HAB, Munich, ETH, manuscrits suisses, etc.
- interfaces spécialisées, par exemple pour un manuscrit particulier (Madame Bovary de Flaubert)
- il existe un certain nombre de travaux sur les critères d'évaluation, mais ils ne seront pas évoqués ici

Gallica : page initiale



- trop de choses
- informations inutiles pour 99 % des gens
- l'utilisateur moyen vient pour une requête précise
- 90% de cette page représente une distraction
- la distraction pourrait être accessible depuis un lien « *la buzz* de Gallica » et cela suffirait

Gallica : résultat d'une recherche

The screenshot shows the Gallica search results page for the query "connaissance des temps". The page features a search bar at the top with the query and a "Go" button. Below the search bar, there are navigation tabs for "All Gallica", "Books", "Manuscripts", "Maps", "Images", "Periodicals", "Sound recordings", and "Scores". The search results are displayed in a list format, with the first result being "Réflexions sur le parti qu'il convient de prendre pour la continuation de la Connaissance des Temps - Paris 1821-1852". The page also includes a sidebar with filters for "Type of document", "Author", and "Publication date".

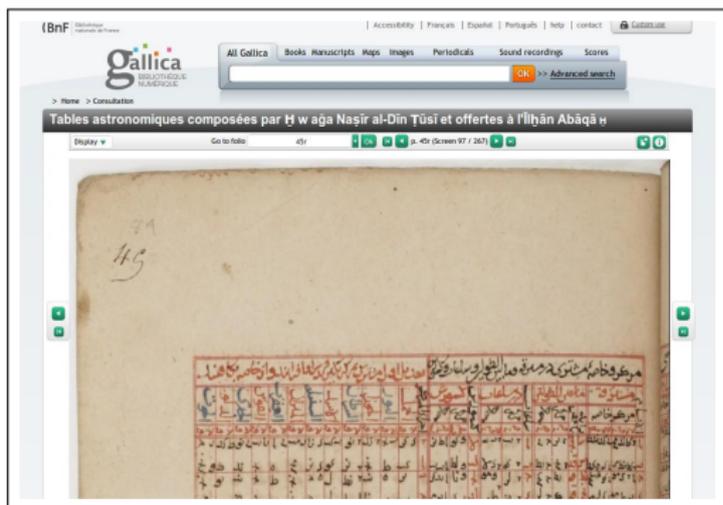
- informations peu utiles à gauche
- trop peu de vignettes et vignettes neutres
- trop de texte
- pas très efficace
- recherche probablement incomplète

Gallica : navigation (suite)

The screenshot shows the Gallica website interface. At the top, there is a navigation bar with the Gallica logo and various menu items: All Gallica, Books, Manuscripts, Maps, Images, Periodicals, Sound recordings, Scores. Below this is a search bar with a 'OK' button and a link to 'Advanced search'. The main content area displays the title 'Tabule astronomice Alfonso regis ([Reprod.]) Alfonso X (roi de Castille et de León : 1221-1284)'. Below the title, there is a 'Display' dropdown menu and a 'Go page' field. A list of 'Page' thumbnails is visible on the left side. The main preview area shows a manuscript page with a table of astronomical data. The table has several columns and rows of numbers and text, including the heading 'Abula de quadrans Antero.'.

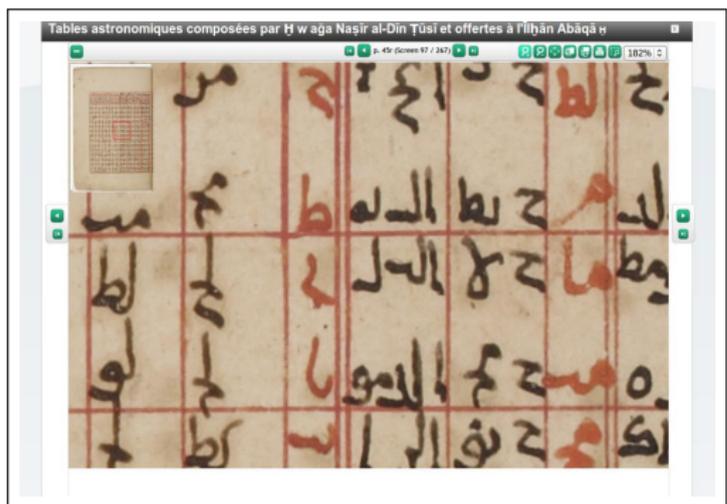
- la liste des pages n'indique pas bien celle où l'on est (elles s'appellent ici toutes NP = non paginé)
- on a la possibilité d'afficher des vignettes, mais indirectement par « display »

Gallica : navigation (suite)



- cas d'un manuscrit
- on ne peut zoomer qu'indirectement en entrant dans le menu « display »

Gallica : navigation (suite)



- le zoom dans le manuscrit via une applet

Contentdm

- plate-forme développée par l'OCLC, organisme fournissant un support informatique aux bibliothèques
- utilisée par Strasbourg, Linda Hall, Warwick, etc.
- <http://docnum.u-strasbg.fr>
- <http://lhldigital.lindahall.org>
- <http://contentdm.warwick.ac.uk/index2.php>

Contentdm : affichage d'une collection

ent Bibliothèque numérique Service Commun de la Université de Strasbourg

Documentallog

Sélectionnez une collection dans la liste suivante : **Mathématiques**

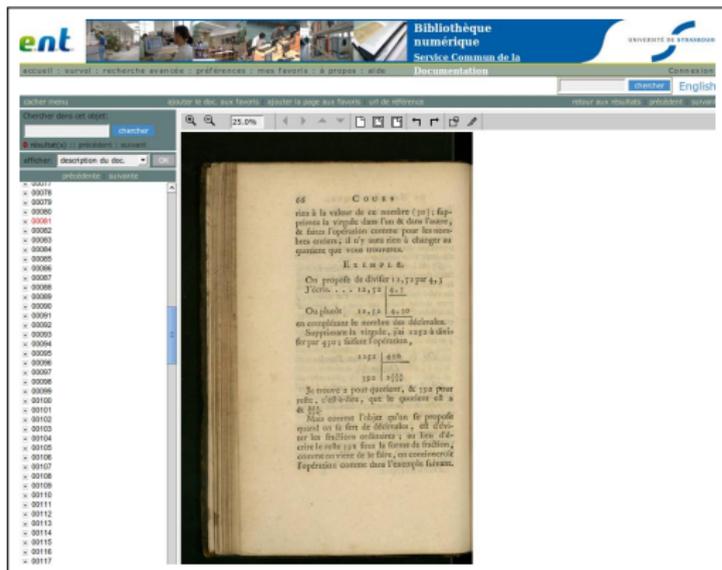
résultats 1-20 sur 126 document(s) page 1 sur 7 : (<< 1 2 3 4 5 6 7 >>) :: précédente | suivant

tout sélectionner | tout effacer | ajouter aux favoris

	Image:	Titre:	Auteur(x):	Collaborateur(x):
<input type="checkbox"/>		Abhandlungen aus der reinen Mathematik.	VANDERMONDE, Alexandre-Théophile	
<input type="checkbox"/>		Algèbre	NETTO, Eugén	
<input type="checkbox"/>		Algebraische Analysis	HATTENDORFF, Karl	
<input type="checkbox"/>		Allgemeine Arithmetik : I. Teil (Zahlenlehre), Vorlesungen vom Winter Semester 1885/90	KRONECKER, Leopold	

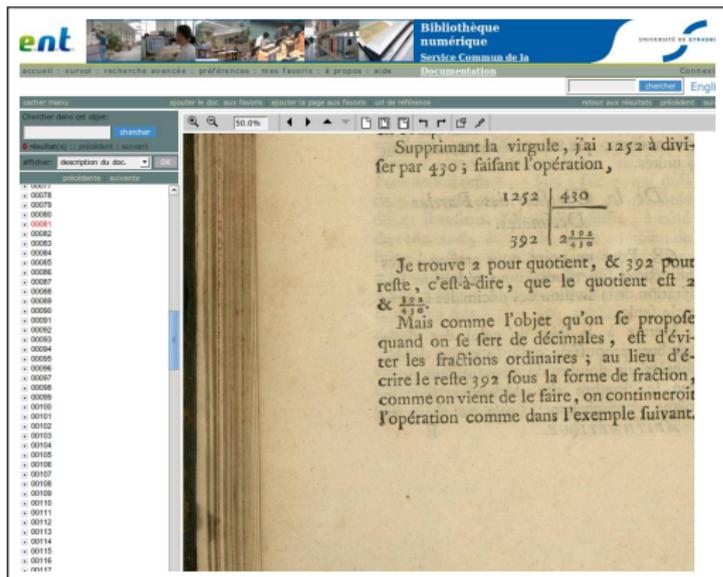
- affichage d'une sous-collection
- les collections devraient être remplacées par une recherche par sujet
- énormément d'espace vide
- les auteurs sont trop loin des titres
- le changement de page en haut à droite est souvent hors de la fenêtre qui n'est pas assez dynamique

Contentdm : vue d'un document



- vue d'un document
- on peut faire une recherche sur l'OCR du document, mais l'OCR n'a pas été vérifié
- le nombre de clics à faire est trop grand (souvent, on doit appuyer sur OK)
- la page est téléchargeable en faible résolution, mais pas facilement en zoom ;
- mais il n'y a pas de téléchargement de tout le document en PDF

Contentdm : le mode zoom (1)



- ici, le bouton « suivant » n'est que partiellement visible
- le déplacement sur la totalité de la page n'est plus possible en agissant uniquement sur l'image! ce problème est presque exclusif de Contentdm.
- on peut contourner le problème du zoom avec les flèches noires pleines, mais ce n'est pas intuitif

Contentdm : le mode zoom (2)



- ce n'est pas intuitif car on a déjà un déplacement partiel (et naturel) de l'image en tirant la souris ;
- ce dont l'utilisateur a besoin, c'est d'agir directement au niveau du texte qu'il voit, pas ailleurs ;
- le déplacement par les flèches introduit en outre un saut trop grand, et n'est donc pas fluide, et donc pas confortable ;
- il y a donc ici un problème évident de conception d'interface ;

Contentdm : le mode zoom (3)

- Certaines fonctionnalités annoncées ne fonctionnent pas du tout, ou sont dépendantes de certaines plate-formes.
- Par exemple, le bouton en haut à droite de l'interface sert à « découper l'image dans une nouvelle fenêtre ».
- Mais dans cette fenêtre, on peut sélectionner une partie, puis paraît-il sélectionner une opération avec le bouton droit de la souris. Or, ceci ne fonctionne pas partout !
- De manière générale, le zoom ne peut pas être positionné assez finement, même avec la petite vignette de visualisation du cadre de zoom (seules certaines positions sont autorisées).

Contentdm

- il est facile de trouver encore d'autres problèmes, tant cette interface semble avoir été mal conçue ;
- les problèmes observés à Strasbourg se retrouvent sur les sites de Linda Hall et Warwick
- des améliorations sont peut-être possibles, mais elles ne sont pas faites
- les (quelques) utilisateurs se sont probablement habitués au handicap qu'on leur impose
- une interface telle que Contentdm est dissuasive

Visual Library

- développé en Allemagne
- utilisé dans le projet suisse e-rara
- utilisé à Düsseldorf

e-rara (Visual Library) : résultat d'une recherche

The screenshot shows the e-rara website interface. At the top, there is a search bar with the text 'Last results' and 'Search' next to a magnifying glass icon. Below the search bar, there are navigation links: 'All libraries', 'Bibliothèque de Genève', 'ETH-Sbibothek Zürich', 'UB Basel', 'UB Bern', 'ZB Zürich', and 'Other libraries'. On the left side, there is a 'Collections' menu with options like 'All collections (3814)', '15th century Swiss prints (11)', and '16th century Swiss prints (2258)'. The main content area displays '19 results' for the search term 'sinuum'. Each result includes a small thumbnail image of the book cover, the title, author, and publication information. The first result is 'Instrumentum sinuum, seu primi mobilis Apianus, Petrus Norimbergae : apud Iohann. Petresium, 1541'. The second result is 'Adriani Vlacq Tabulae sinuum, tangentium, et secantium: logarithmi sinuum, tangentium, et numerorum ab unitate ad 10000 Hentsch, Johann Jacob ; Vlacq, Adrian Francofurti : Impensis Joh. Friedr. Fleischli, 1757'. The third result is 'Doctrina triangulorum sive trigonometria die Lehre von Messung der Trianglen sambt dem Gebrauch der Tabularum Sinuum, Tangentium & Secantium Meyer, Georg Friedrich ; Meyer, Jacob Bessel : Geneth, 1678'. The fourth result is 'Nicolai Raymani Ursi ... Fundamentum astronomicum, id est nova doctrina sinuum et triangulorum eaque absolutissima et perfectissima, eiusque usus in astronomica [...] Ursus, Nicolaus Reimarus Argentorati [Strassburg] : excudebat Bernhardus Jobst, 1588'. The fifth result is 'Ioannis de Monteregio ... Tabulae directionum projectionumque, non tam astrologiae iudicialia, quam tabulis Instrumentisque innumeris fabricandis utiles [...] eiusdem Regionontani tabula sinuum per singula minuta extensa, universam sphaericorum triangulorum scientiam complectens Regionontanus, Johannes'.

- interface simple
- page de recherche sur un terme (ici « sinuum »)
- cinq vignettes et une courte description (mais où l'auteur n'est pas toujours bien en évidence)
- vignettes ni trop grandes, ni trop petites
- on clique sur la vignette ou le titre pour atteindre l'ouvrage

e-rara (Visual Library) : page principale d'un document



The screenshot shows the e-rara (Visual Library) interface. At the top, there is a header with the e-rara logo on the left and the ETH-Bibliothek logo on the right. Below the header is a navigation bar with tabs for 'Title view', 'Content', 'Thumbnails', and 'Page view'. A search bar is located in the center of the header. The main content area is divided into two columns. The left column contains a thumbnail of a book cover with the title 'NICOLAI COPERNICI TORINENSIS DE REVOLUTIONIBUS ORBIVM COELESTIVM, LIBRI VI'. The right column contains a table of bibliographic metadata.

Bibliographic Metadata	
Title	Nicolai Copernici Torinensis de revolutionibus orbium coelestium, Libri VI
Author, Contributor	Copernicus, Nicolaus ^W
Imprint	Nürnberg : apud Joh. Petreium, 1543
Description	[6]+196 Bl., [1] Bl. 1f. (Holzschnitte im Text) 29 cm
Bibliographical reference	VD 16 K 2099
Language	Latin
Owner of original copy	ETH-Bibliothek Zürich, BIL 1370 ^W
Persistent Identifier	10.3931/e-rara-420
Links	
Entre Book as PDF	[1] Nicolai Copernici Torinensis de revolutionibus orbium coelestium, Libri VI [109.07 rna]
Collection	
Other collections	Astronomy (ETH-Bibliothek)
Content	
Table of Contents	

- accès aux méta-données, y compris la localisation de l'original
- accès vers la vue « Page » ou « Thumbnails »
- téléchargement immédiat (mais dégradé)
- d'autres possibilités comme le parcours par chapitres sont aussi possibles

e-rara (Visual Library) : vue d'une page



- on peut avancer dans le document, choisir une page quelconque, ou zoomer (sans restriction de zone) à partir des boutons du haut
- les boutons sont trop petits et trop proches
- deux modes de zoom début 2011 (ici), mais réduits à un seul en 2012 (rendant plus difficile le téléchargement d'une bonne image)
- rotations possibles
- navigation hiérarchique

e-rara (Visual Library) : vignettes



- le mode vignettes affiche un certain nombre de vignettes du document ;
- on peut passer d'une page de vignettes à la suivante ;
- on peut aussi accéder à un intervalle de vignette par un menu ;
- mais un parcours rapide (type Google) est impossible

e-rara (Visual Library) : qualité de texte en PDF

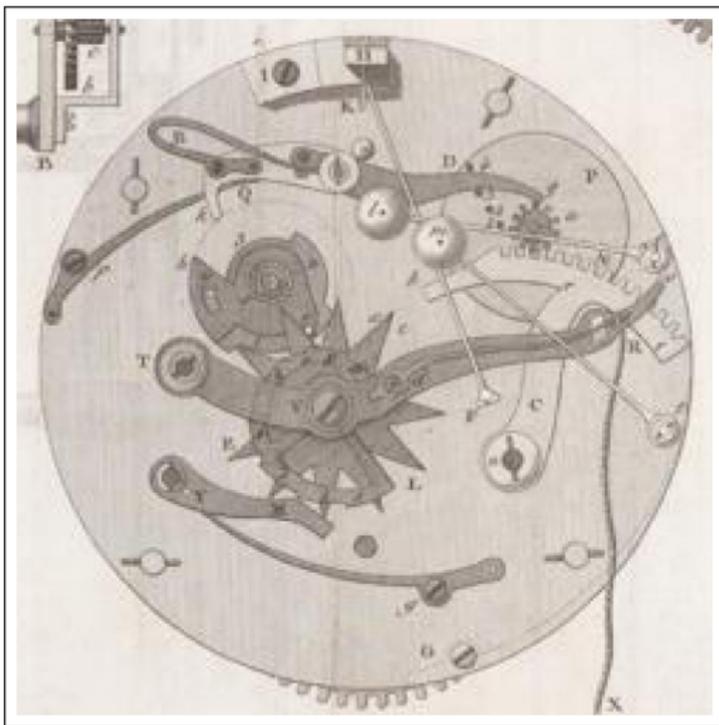
16 HISTOIRE DE LA MESURE DU TEMPS.

une description^a de ces différens instrumens. Nous allons en rapporter quelques détails, en distinguant ce qui semble dû à cette École, de ce qui appartient à des temps antérieurs.»

Les premières horloges d'eau ont été simples et même grossières. On aura d'abord voulu mesurer le temps par l'eau écoulée d'un vase : mais on n'aura pas tardé à s'apercevoir que les quantités d'eau n'étoient pas proportionnelles au temps ; et, après avoir reconnu que l'erreur naissoit de la chute inégale de l'eau, on aura ensuite cherché à y remédier en employant, au contraire, le temps de l'immersion des corps dans l'eau. Le petit bateau des Indiens, percé d'un trou, qui surnage d'abord et s'enfonce au bout d'un certain temps fixé par l'expérience, a peut-être été, dans ce genre, le premier degré de perfection des clepsydes.

- Berthoud : *Histoire de la mesure du temps par les horloges* (1802)
- le format PDF pour le texte est acceptable

e-rara (Visual Library) : qualité de figures en PDF



- Berthoud : *Histoire de la mesure du temps par les horloges* (1802)
- le format PDF pour la figure est extrêmement mauvais
- par ailleurs, les images ne peuvent plus être téléchargées directement avec une bonne qualité
- le format interne est très détaillé, mais n'est pas téléchargeable, sauf en mosaïque

Dresden : page principale d'un document

The screenshot shows the main page of a document record in the SLUB Dresden digital library. The interface is organized into several sections:

- Header:** "SLUB Dresden" logo and navigation tabs for "TITELKARTEN" and "NACHWEISEN".
- Search:** A search bar with a "Suchen" button and options for "In Bibliothek", "Im SLUB-Katalog", and "In Digitalisat".
- Left Sidebar:**
 - Digitale Bibliothek:** "Digitale Sammlungen" and a note: "Möchten Sie sich hier an, um unsere personalisierten Dienste zu nutzen."
 - Benutzersitzung:** Fields for "Benutzername" and "Passwort" with an "Anmelden" button. Below are checkboxes for "Anmelden bleiben?" and "Nach nicht registriert?".
- Main Content Area:**
 - Titel:** "Prutenicae tabulae coelestium motuum"
 - Strukturtyp:** Monographie
 - Autor:** Erasmus Reinhold
 - Sprache:** Latein
 - Digitale Fassung:** Includes a "zur Werkansicht" button.
 - Persistente URL:** <http://digital.slub-dresden.de/id278475426>
 - PPN:** 278475426
 - URN:** urn:nbn:de:hbz:14-dp-id2784754263
 - Nachweise:** "SLUB-Katalog", "Südwestdeutscher Bibliotheksverbund"
 - Sammlung:** "Technikgeschichte"
 - Digitalisierer:** "SLUB Dresden"
 - Förderer:** "Deutsche Forschungsgemeinschaft"
 - Original:**
 - PPN:** 626178095
 - Nachweise:** "SLUB-Katalog", "Südwestdeutscher Bibliotheksverbund"
 - Besitzer:** "SLUB Dresden"
 - Signatur:** "Astron.191.d7.misc.1"
 - Erscheinungsort:** "Wittenbergae"
 - Erscheinungsjahr:** "1585"
 - Umfang:** "[8], 68, [1], [1], 142 Bl., [2] gef. Bl."
- Right Sidebar:**
 - Kontakt:** "Fragen Sie uns!", "SLUB Dresden", "Postanschrift: 93954 Dresden", "Besuchersadresse: Zölchstr. Ring 18, 93069 Dresden"
 - Werkzeuge:** "Nutzungsbedingungen", "Digitale Kollektionen erstellen...", "Gesamtwort als PDF herunterladen...", "Leserzeichen setzen...", "Im DFG-Virenet betrieblen..."
- Image:** A photograph of a modern library interior with bookshelves and study tables.

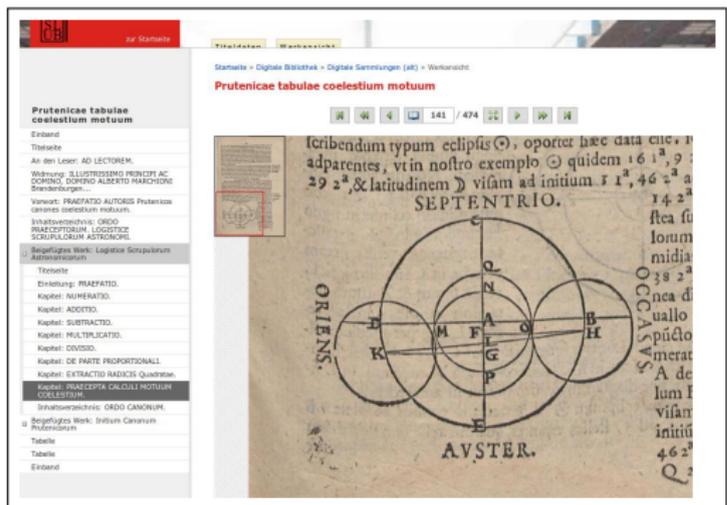
- aucune vignette du document
- beaucoup de métadonnées
- seules quelques unes de ces métadonnées sont essentielles

Dresden : vue du document

The screenshot shows a digital library interface for the document 'Prutenicae tabulae coelestium motuum'. The interface is in German and includes a header with the library name 'Sächsische Landesbibliothek - Staat- und Universitätsbibliothek Dresden' and a search bar. The main content area displays the title 'Prutenicae tabulae coelestium motuum' and a grid of 12 thumbnail images of manuscript pages, labeled 'Bild 1 von 474' through 'Bild 8 von 474'. A navigation bar above the thumbnails shows '1 / 24' and navigation arrows. The left sidebar contains a table of contents with sections like 'Einband', 'Titelblatt', 'An des Lesers: AD LECTOREM', 'Widmung: ILLUSTRATIONI PRINCIPIS AC DOMINI GONRADO ALBERTO PRINCIPIS Brandenburgerum...', 'Vorwort: PRASFATIO AUCTORIS Prutenicae coelestium coelestium motuum.', 'Inhaltsverzeichnis: DREDO PRACEPIORUM LOGISTICE SCRIPULORUM ASTRONOMIC.', 'Beigefügtes Werk: Institutum Sphaularium Astronomicorum', 'Beigefügtes Werk: Institutum Canonum Prutenicarum', 'Tabelle', and 'Einband'.

- par défaut le document est présenté en mode « miniatures »
- on peut circuler d'un groupe de miniatures au suivant
- téléchargement PDF possible

Dresden : zoom sur une page



- détail d'une page
- on peut circuler naturellement et, de plus,
- on peut zoomer avec la molette, mais comme c'est l'un des seuls sites qui le propose, cela pose des problèmes en changeant de site

Dresden : version beta

[http://www.slub-dresden.de/index.php?id=5363&tx_dlf\[id\]=18717](http://www.slub-dresden.de/index.php?id=5363&tx_dlf[id]=18717)

The screenshot displays the SLUB Dresden website interface. At the top, the logo and name 'Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden' are visible. The main content area shows a manuscript page titled 'Divisores numerorum decies centena millia non excedentium, c' by Aster Kalk, Jakob Philipp, from 1825. The page features a large table of numbers arranged in columns and rows. On the left, there is a navigation menu with options like 'Inhaltsverzeichnis', 'Titelblatt', and 'Einführung'. On the right, there is a 'Werkzeuge' (Tools) section with options for 'einfache Ansicht', 'An den Funktionen des Werkzeugkastens arbeiten', and 'wie danach noch!'. The bottom of the page has a footer with contact information and a 'Digitaleingangsbereich' link.

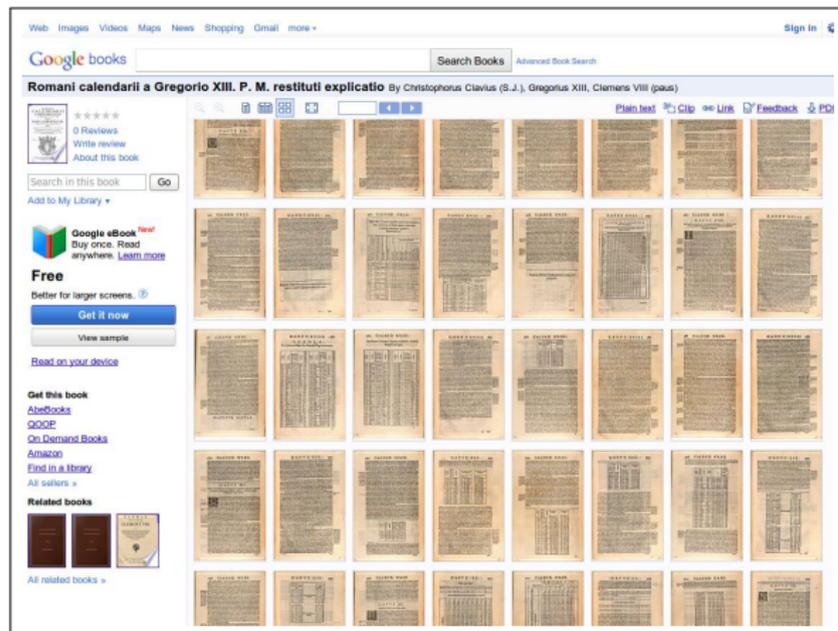
- nouvelle version de l'interface
- dans cette version, le téléchargement PDF n'était plus possible en mars 2011, mais cela est apparemment remis en fonction

Google books

The screenshot shows the Google Books search results page for the book "Canon triangulorum emendatissimus" by Bartholomaeus Pitiscus. The page includes a search bar, a "Search inside" button, and a "Preview this book" link. Below the book cover, there is a section for "What people are saying" and a "Related books" section with five book covers. On the right side, there are sections for "Add to My Library", "Free" (with a "Get it now" button), "Read on your device", and "Get this book" (with links to various retailers like AbeBooks, Google, etc.). At the bottom, there is a "Contents" section showing "Section 1" and "Section 13".

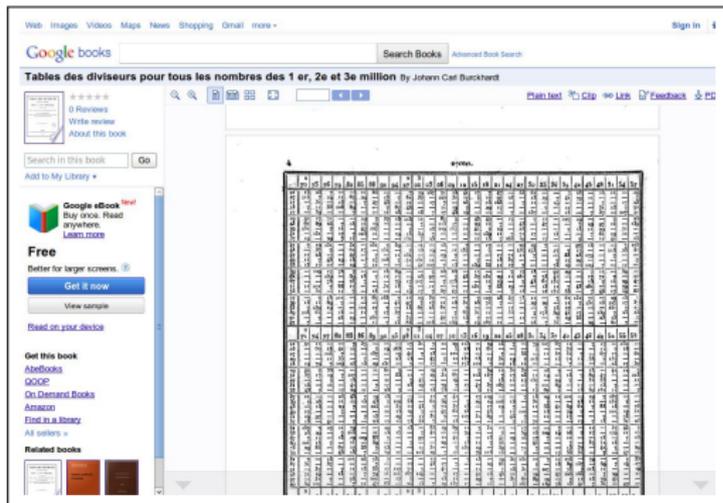
- on ne voit pas la première page du livre
- on ne voit même pas du premier coup qu'il est téléchargeable
- on a même l'impression qu'on ne peut qu'acheter ce livre

Google books : vignettes



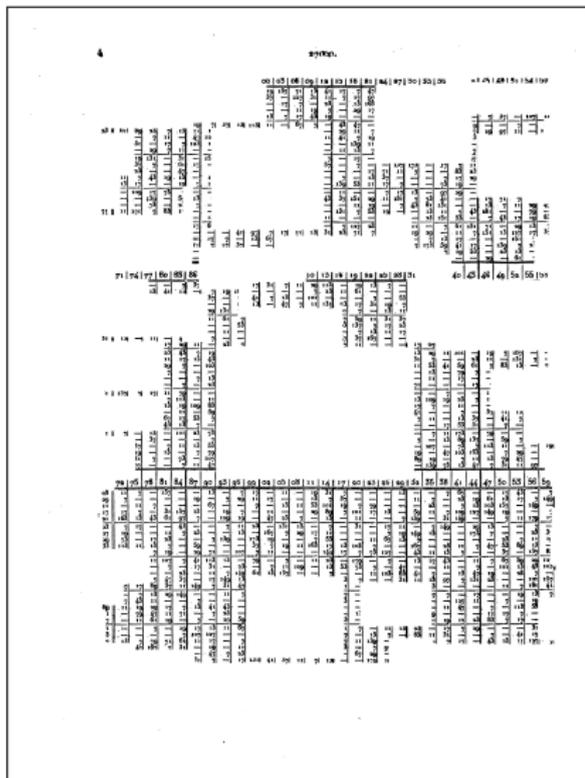
- vignettes
- sélectionnées par un bouton
- défilement par l'ascenseur, ce qui est pratique et inexistant dans les autres interfaces

Google books : détail d'une page



- la page paraît normale
- l'interface est assez simple
- elle privilégie le déplacement avec l'ascenseur ou la molette
- une partie de la fenêtre de gauche est de nature commerciale
- l'année du livre n'est pas indiquée ici

Google books : représentation interne d'une page



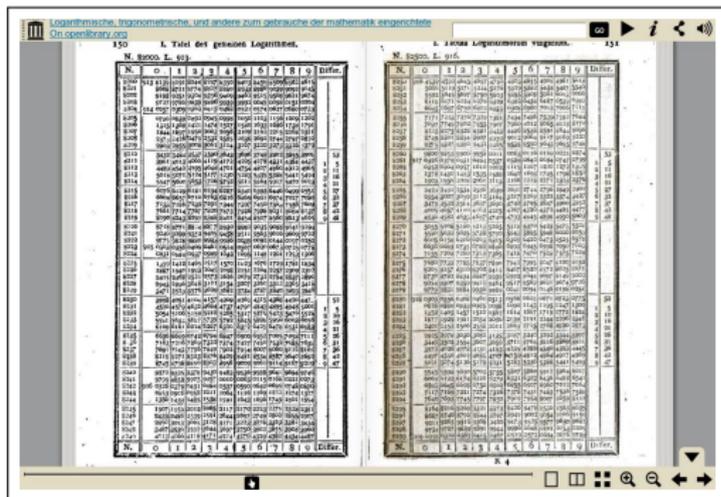
- la page est en fait trouée
- dans le PDF, elle est stockée sous forme de couches (format JPEG2000)
- le découpage en couches est souvent nettement visible dans l'original
- cela nuit à la qualité de l'image
- on peut penser que l'interface d'affichage de Google est imparfaite

Archive.org

The screenshot shows the American Libraries website interface. At the top, there are navigation links for Home, Canadian Libraries, Universal Library, Community Texts, Project Gutenberg, Children's Library, Biodiversity Heritage Library, and Additional Collections. A search bar is present with the text 'American Libraries' and a search button. Below the search bar, there is a breadcrumb trail: 'Book and Text Archive > American Libraries > Logarithmische, trigonometrische, und andere zum gebrauch der mathematik eingerichtete tafeln und formeln (1783)'. The main content area is divided into two columns. The left column, titled 'View the book', contains a small image of the book cover and a list of download options: Read Online (496 pp, 121.9 Mb), PDF (Google.com) (496 pp), ePub (496 pp), DjVu (496 pp), and Full Text (1.4 Mb, 351.4 Mb, 126.9 Kb). Below this is a 'Resources' section with a 'Bookmark' button. The right column contains the book's metadata: Author: 'Johann Georg Trithem von 1734-1802', Subject: 'Logarithms', Publisher: 'Wien, Gedruckt bey Johann Thomae, neben von Trattner', Possible copyright status: 'NOT_IN_COPYRIGHT', Language: 'German', Digitizing sponsor: 'Google', Book from the collections of: 'University of Michigan', and Collection: 'arctools'. A link indicates 'This book has an editable web page on Open Library'. Below the metadata is a 'Description' section stating 'Book digitized by Google from the library of University of Michigan and uploaded to the Internet Archive by user jpb.' and 'Bibliography: p. ii'. There is a 'Reviews' section with a 'Be the first to write a review' button and 'Downloaded 27 times'. At the bottom, there is a 'Selected metadata' section with fields for Google-id, Identifier, Scanner, Source, and Mediatype.

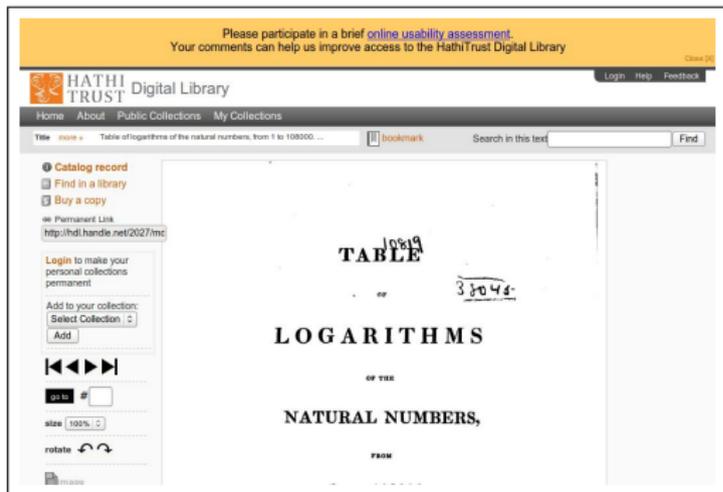
- une partie des livres d'archive.org proviennent indirectement de Google books, mais le format PDF n'est pas le même
- archive permet aussi un défilement par un feuilletage virtuel
- qualité analogue à Google
- lien vers le projet « Open Library »
- le livre est quelquefois téléchargeable sur Archive, mais pas sur Google

Archive.org : détail d'une page



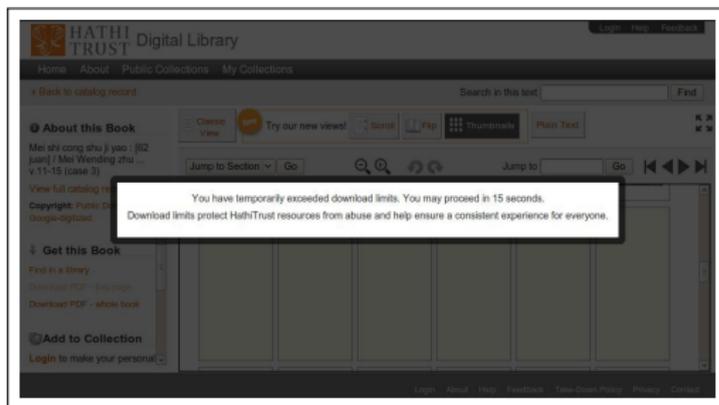
- le déplacement rapide peut se faire avec la barre du bas
- artéfacts colorés
- certaines parties paraissent plus nettes que d'autres en raison d'une structuration par couches des pages

Hathi Trust (1)



- un autre nouveau venu
- certains livres proviennent indirectement de Google (en fait des reversements de numérisations Google)
- les restrictions peuvent être différentes

Hathi Trust (2)



À moins de faire partie d'institutions privilégiées, certains documents (ici de 1761) ont de telles restrictions d'accès qu'il n'est même pas possible de parcourir les vignettes sans régulièrement faire face à un écran pour ralentir la recherche.

Hathi Trust (3)

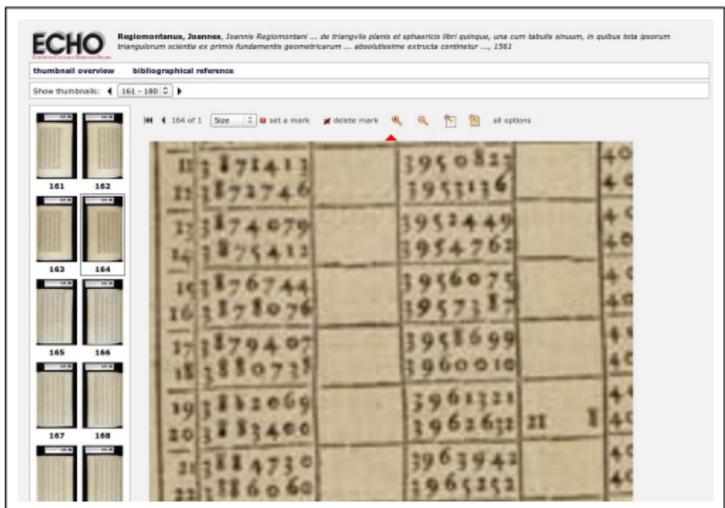
- l'interface de Hathi manque tellement de convivialité qu'elle en est repoussante ;
- la consultation de documents pour la recherche est très difficile, en raison des restrictions ;
- l'interface comptabilise apparemment les accès aux pages, mais pondère identiquement les vignettes et les pleines vues ;
- or, pour rechercher rapidement une partie (par exemple une image), les vignettes sont indispensables, et ces vignettes consomment alors tout le crédit ;
- la vue normale ne permet pas un défilement rapide ;
- Hathi semble être un exemple de site peut-être très fourni, mais probablement quasiment inutilisé, car inutilisable.

Berlin



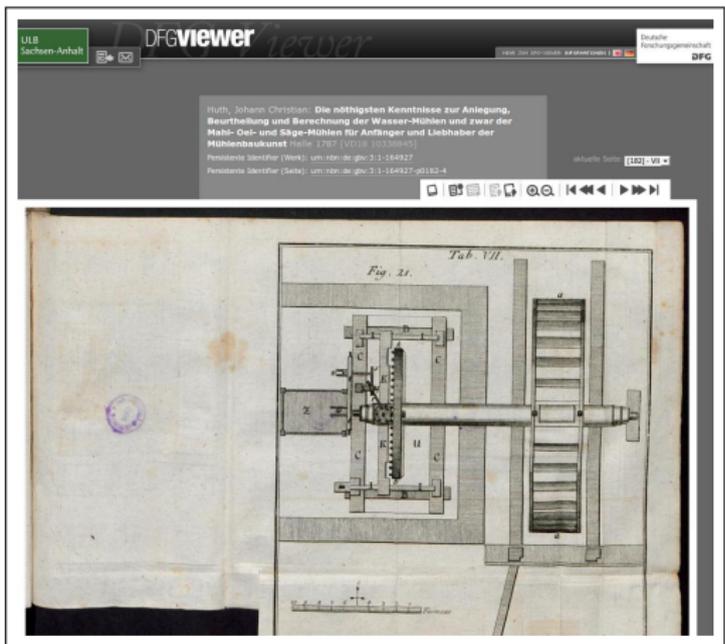
- téléchargement apparemment impossible
- le document provient de HAB (bibliothèque de Wolfenbüttel)

Berlin : le zoom



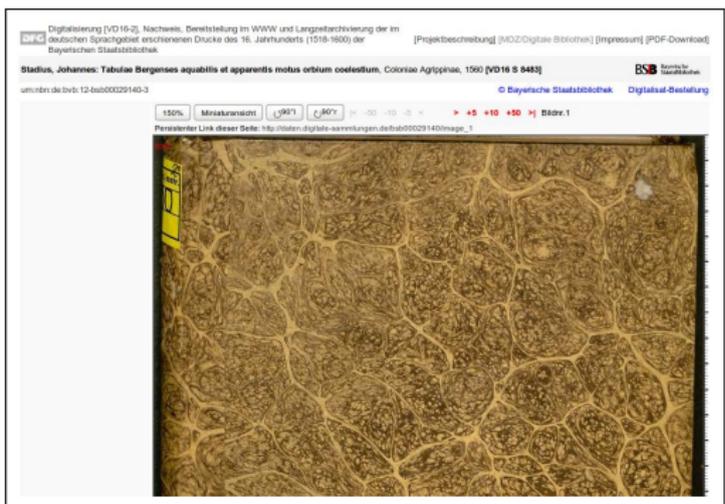
- la résolution est grandement insuffisante
- les chiffres devraient être plus nets
- est-ce la célèbre bibliothèque de Wolfenbüttel qui n'a pas été capable de numériser adéquatement ?
- l'interface de zoom offre un déplacement très **incommode**, pire que celui de Contentdm

ULB Sachsen-Anhalt



- exemple où les planches ont été numérisées (pliées et dépliées)
- mais pas très bien alignées
- ...

Munich : page principale d'un document



- vue sur la couverture en premier (pas très utile)
- téléchargement possible

Munich : interface des numérisations de Google

Munich met aussi en ligne des numérisations de Google, mais avec une interface différente :

BSB Bayerische Staatsbibliothek digital Katalog (OPAC) | Impressum | English

Dittmann, C.: Koordinaten- und Tangenten-Tafeln (1859) Bibmhr

Seite 73

73

Grad.										Minuten
5	6		7		8		9			
Abcisse	Ordinate									
409 5760	344 1458	491 4912	401 5034	373 4064	458 3611	453 3216	518 2767	737 2308	60	
4926	2888	3911	6702	2996	459 0517	1881	4332	8066	59	
4091	4317	2909	8370	1727	2423	0545	6476	736 9363	58	
3266	5746	1907	400 0087	0538	4328	654 9209	8619	7800	57	
3420	7175	6904	1704	572 9288	6233	7872	517 0762	6396	56	
1584	8603	490 9901	3370	8218	8127	6535	2904	4852	55	
0748	845 0031	8898	5036	7047	460 0041	5197	5046	3347	54	
408 9912	1459	7894	6692	0876	1945	3859	7188	1841	53	
5075	2586	6590	8667	4765	3848	2670	9329	0335	52	
8238	4314	5885	403 0033	3533	5752	1190	518 1471	738 8828	51	
7400	5740	4880	1097	2360	7654	653 9810	3611	7320	50	
6562	7167	3875	3862	1187	9566	8510	5751	5812	49	
5724	8508	2869	5026	0014	461 1438	7109	7890	4304	48	
4886	346 0020	1863	6630	571 8840	3360	5817	519 0030	2794	47	
4047	1445	0856	8553	7665	5200	4475	3168	1284	46	
3407	2871	489 9849	404 0016	6490	7163	3132	4306	734 9733	45	
2368	4296	8841	1678	0315	5081	1788	6444	8502	44	
1528	5721	7853	3341	4139	409 0961	0444	8581	6750	43	
0688	7145	6825	5003	2963	2860	652 9100	520 0718	5208	42	
407 9847	8570	3817	6665	1786	4760	7756	2855	3723	41	
9006	9993	4807	8326	0609	6658	6410	4990	2211	40	
8165	317 1417	3798	9287	570 8431	8556	3064	7126	0607	39	
7323	2840	2788	1047	3252	468 0454	3717	9261	733 9182	38	
6481	4264	1777	3308	7074	2352	9270	521 1396	7666	37	
5639	5687	0766	4968	5894	4249	1022	3530	6150	36	
4796	7109	488 9755	6628	4715	6146	651 9674	5664	4623	35	
3963	8531	8745	8287	3534	8042	8324	7797	3115	34	
3109	9953	7241	9946	3338	9936	6975	9389	1507	33	

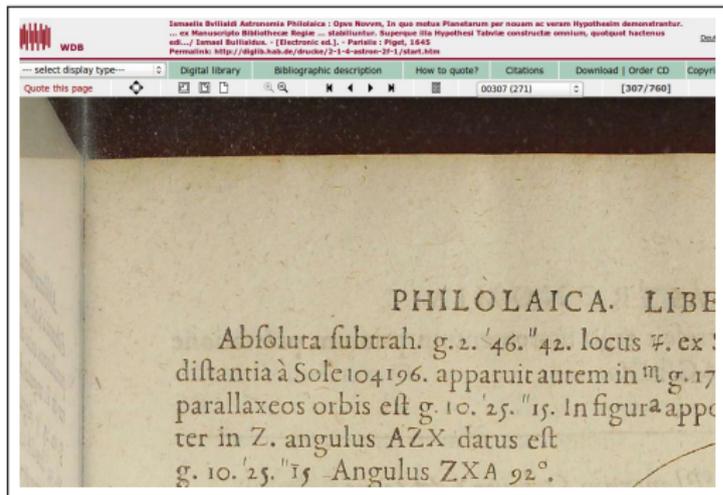
- on conserve les boutons d'avance rapide, mais l'interface reste assez limitée

Goettingen

The screenshot displays the GDZ website interface. At the top left is the GDZ logo and the text 'Göttinger Digitalisierungszentrum'. To the right, it identifies the document as 'Teutsche Algebra oder Algebraische Rechenkunst' by Leonhard Euler, a monograph from 1770. A navigation bar includes buttons for 'Seitenansicht', 'DFO-Viewer', 'Inhaltsverzeichnis', 'Bibliographische Info', and 'PDF Download'. Below this is a digital viewer window showing a manuscript page with the title 'Teutsche Algebra/ Oder Algebraische Rechenkunst/ zusamt ihrem Gebrauch: Vesehend' and a list of three items: 1. In Auflösung vorvornher Mathematischer Aufgaben, 2. In Verhandlung allerhand Algebraischer Equationen, 3. In Erfindung unterschiedlicher nutzlicher Theorematum. The left sidebar contains various utility links like 'Dokumenten-Server Browser', 'Schreibtaische', and 'Serviceleistungen'. A small thumbnail of an open book is visible at the bottom left of the sidebar.

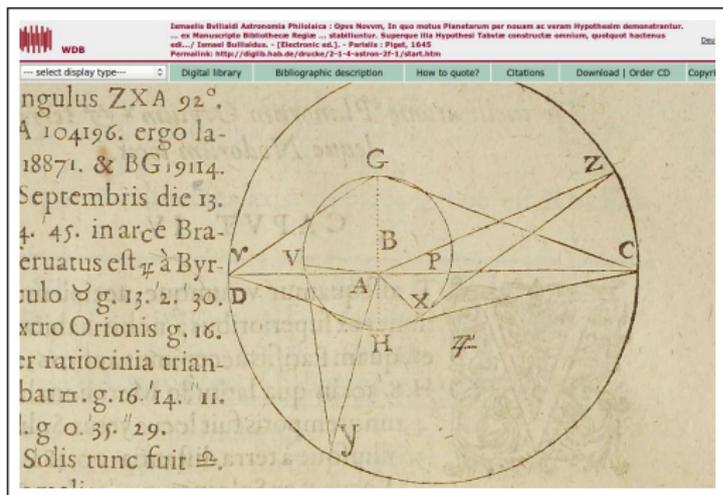
- ici, comme sur d'autres interfaces, il serait utile d'avoir des boutons d'avance rapide

HAB (Wolfenbüttel)



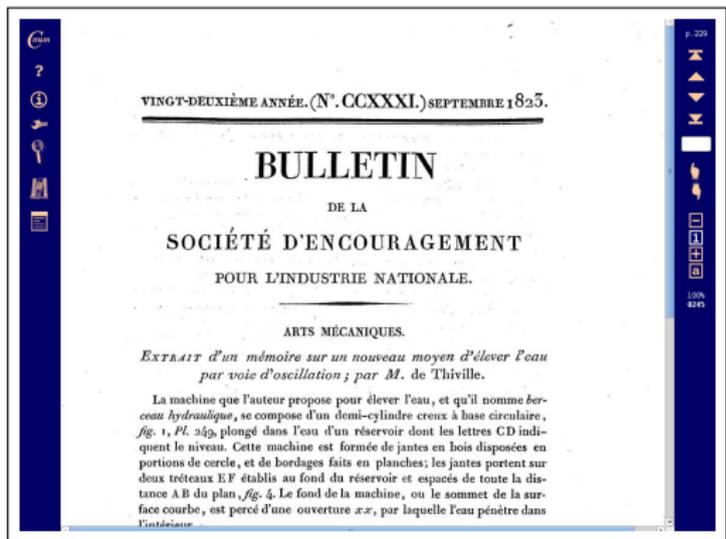
- vue en zoom en haut de page
- ici, la résolution est excellente
- dans le document vu précédemment, il n'y a en fait que la version basse résolution

HAB (Wolfenbüttel)



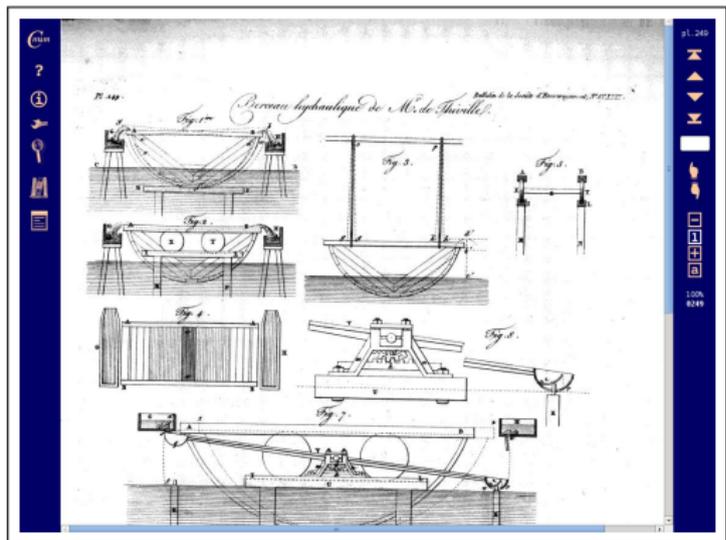
- vue en zoom
- en descendant, on n'a plus accès aux boutons de navigation
- ce problème est fréquent
- Gallica a essayé de le résoudre, mais la solution employée est imparfaite

CNUM (CNAM)



- pages d'un document numérisé par le CNUM <http://cnum.cnam.fr>
- interface très limitée
- les « mains » permettent un parcours séparé des planches
- pas de page de vignettes, ni de parcours rapide par ascenseur
- un PDF complet est téléchargeable sur la page de sommaire du document

CNUM (CNAM)



- vue d'une planche
- celle-ci n'est pas alignée et sa résolution est insuffisante

IMSS (Florence)

Viète, François *Franc. Vietæ Libellorum supplicum in regia magistri insignique mathematici varia opera mathematica. In quibus tractatus...*

opere presentata | BVB

INDICE

- Copertina
- Frontispizio
- Stantibus Operum Canon Mathematicus seu ad Triangula
- Ad Lectorem
- Occhioletto: Canon Mathematicus, seu ad Triangula
- Canon Mathematicus, seu ad Triangula
- Canonis Triangulorum Laterum Rationalium
- Capita Inspectionum
- Francisci Vietæ, Universalium Inspectionum ad Canonem Mathematicum, Liber Singularis
- Speculæ Uranicæ
- Copertina

Abita schema

Zoom

Lente

Ruota

Cerca

Scarica

Page: [000]

© 2018 PIRIS - P220 64 04010 1 - 91122 PIRIS00 - ITA02A 14 - +39 019 2483111 - info@numisglobe.it - Urbino aggiornata: 27-08-2010

- téléchargement possible, mais seulement en basse résolution (insuffisante)
- pages N&B et tatouées
- le zoom fait ouvrir une fenêtre séparée
- les boutons de zoom et autres sont à gauche, trop loin de l'image
- déplacement dans le texte avec la barre du bas
- trop grand éparpillement des accès

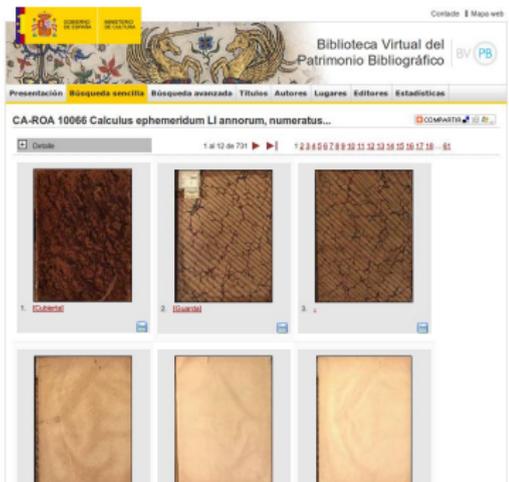
Séville

The screenshot shows the 'Fondo Antiguo' website from Universidad de Sevilla. The main content area displays the title 'Canon trigonometricus : continens logarithmos, sinuum, et tangentium, ad singula scrupula, totius semicirculi'. Below the title is a thumbnail of the book's title page, which reads: 'CANON TRIGONOMETRICVS. CONTINENS LOGARITHMOS, SINVM, ET TANGENTIVM, ad fingula scrupula totius Semicirculi. Ratio Logarithm. 10.00000000. MATRILE. Apud Bernardum à Villa-Diego.'

The interface includes a search bar, a navigation menu on the left with categories like 'Books', 'Illustrations', and 'Historical newspapers', and a right sidebar with metadata sections: 'Authors' (Zaragoza, José (S.I.), 1627-1678), 'Data from the play' (Publication date: 1672, Publication place: Madrid, Signature: A 627/0432), and 'Book's groups' (Folleto natural y otras de carácter científico, Libros del siglo XVII).

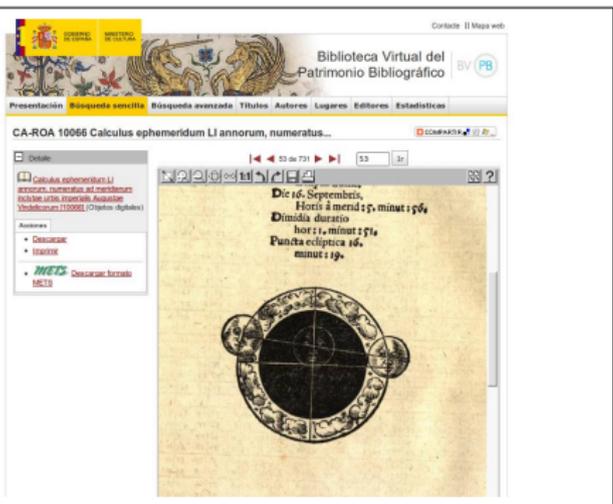
- numérisations N&B

Biblioteca Virtual (Espagne)



- la page d'entrée d'un document donne les miniatures

Biblioteca Virtual (Espagne) : détail d'un document



- détail d'une éclipse de 1559
- on peut utiliser la molette pour zoomer

Heidelberg

RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG

UNIVERSITÄTS-BIBLIOTHEK

Startseite » Elektronische Medien »

HEID | Web-Seiten

Literatursuche und -bestellung

Elektronische Medien

Nutzung und Service

Fachbezogene Informationen

Schulungen

Bibliotheken der Universität

A bis Z

Aktuelles

UB für Einzelige

Konto

Studiengebühren

Sitemap | Kontakt | Layout anpassen | English

Heid. Ha. 3394
Heinrich Rüdinger
Planetenbuch
Heidelberg (Wainheim), 1551-1564

Wissenschaftliche Beschreibung
Sammlung

Persistente URL: <http://digi.ub.uni-heidelberg.de/diglit/heidhs3394>
URN: [urn:nbn:de:hbz:16-diglit-3150](http://nbn-resolving.org/urn:nbn:de:hbz:16-diglit-3150)

Bestellung  [Download](#)
(PDF, 35 MB)



[Sprung zur Seite:](#) (z. B.: 12v, 20r)

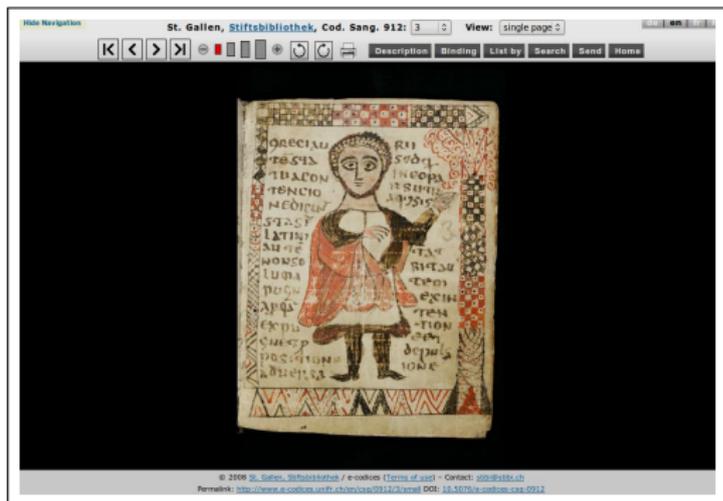
Inhalt

- Einband
- 1ar Titelblatt, Widmung
- 1r Wappen und Devise Ottheinrich
- 1v - 56v Deutscher astronomischer Kalender
- 57r - 75v Johann Mercurius: Geburtsprognostik
- 76r - 93v Petrus Apian: Ufus Almanach
- 94r - 131v Lucas Gauricus: Astrologischer Traktat
- Einband hinten

- on a ici un manuscrit
- la première page donne la structure hiérarchique

Manuscrits suisses : www.e-codices.unifr.ch

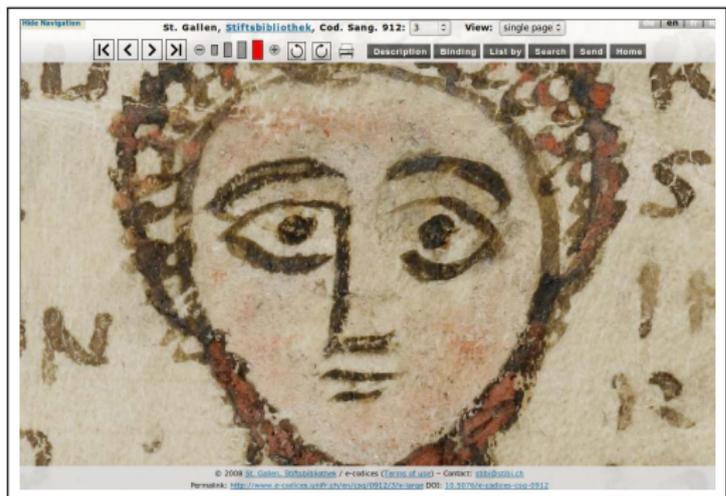
Un manuscrit palimpseste du 5^e siècle



- on ne peut apparemment obtenir un PDF que d'une page à la fois
- on peut aussi récupérer les images séparément

Manuscrits suisses : www.e-codices.unifr.ch

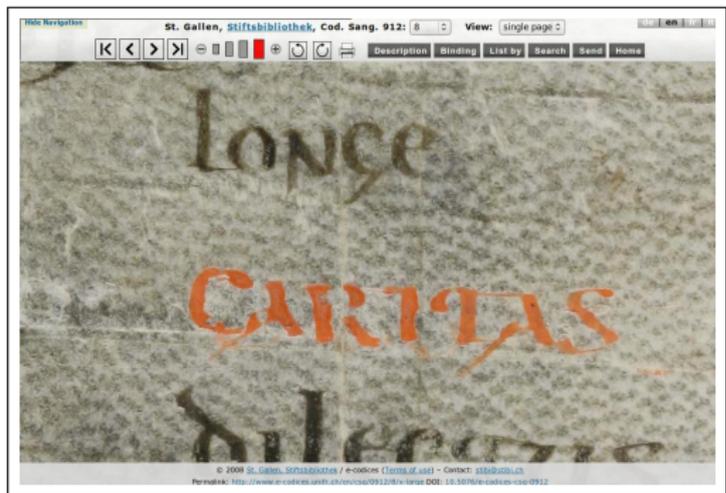
Un manuscrit palimpseste du 5^e siècle



- zoom maximal (en rouge)

Manuscrits suisses : www.e-codices.unifr.ch

Un manuscrit palimpseste du 5^e siècle



- détail d'un mot avec le même zoom

Qualité des interfaces : résumé

- quelquefois affligeant
- soit les utilisateurs ne sont pas pris en compte, soit ce sont des utilisateurs peu exigeants

Une bonne interface doit :

- regrouper des fonctionnalités de base ;
- être ergonomique et autant que possible se rendre invisible ;

Une bonne interface doit être naturelle et discrète, comme une bonne police de caractères.

Il ne suffit pas qu'une interface fournisse toutes les fonctionnalités.

La question du téléchargement

- le téléchargement complet est utile, car tout le monde ne peut pas (ou ne veut pas) lire un ouvrage en ligne, ou sur un autre écran
- la qualité téléchargeable peut ne pas être la qualité réelle
- la qualité téléchargeable doit être suffisante pour une lecture confortable, ce qui n'est pas toujours le cas (à quoi sert une qualité insuffisante ?)
- il n'est pas impossible que certaines bibliothèques aient intentionnellement choisi une interface rendant difficile le téléchargement, mais le manque de transparence de nombreuses bibliothèques ne permet pas de conclure.

La question du téléchargement (2)

Le téléchargement induit une **peur** :

- les bibliothèques sont attachées à leurs livres ;
- elles aimeraient souvent ne pas *donner* complètement les versions numériques ;
- elles aimeraient/aimeront introduire une traçabilité : qui télécharge quoi ?

Ces peurs correspondent à une **vision périmée du fonctionnement des bibliothèques**.

Le métier des bibliothèques devra peu-à-peu être redéfini :

- conservation des originaux
- création de versions numérisées
- enrichissement des versions numérisées

Ce sont les originaux qui restent la propriété naturelle des bibliothèques. Les versions numérisées doivent être distribuées aussi largement que possible.

Critères minimums pour une bonne interface (1)

- pages simples, sans informations inutiles
- titre, auteur, année sur toutes les pages
- fonctionnalité de déplacement habituelles
- déplacements sur des groupes de vignettes
- déplacement sur des vues isolées de 5 en 5, 10 en 10, 50 en 50, 100 en 100 (très utile, mais rarement présent)
- regroupement des fonctionnalités (pas d'éparpillement)
- fonctionnalités moins importantes → page séparée
- éviter les zones blanches (gaspillage)
- interface internationalisée avec au moins l'anglais
- rapidité d'accès

Critères minimums pour une bonne interface (2)

- le zoom ne doit pas restreindre la zone de déplacement, ni nécessiter des opérations spéciales (comme sur contentdm)
- le déplacement (sans restrictions) au sein d'un zoom devrait se faire en tirant l'image
- le nombre de clics doit être minimisé
- les boutons doivent être proches des images
- les vignettes ne doivent être ni trop grandes, ni trop petites
- on doit pouvoir naviguer facilement à zoom constant
- les boutons ne doivent pas être trop près les uns des autres
- téléchargement image et PDF
- réduction de résolution acceptable dans le PDF, mais le problème doit être contournable

Quelles sont les causes de tous ces problèmes ?

Le métier des bibliothèques a changé (1)

- jusqu'à peu, les bibliothèques créaient peu : les livres n'étaient pas fabriqués par les bibliothèques ;
- la bibliothèque était un intermédiaire entre l'éditeur et le lecteur ;
- la création se limitait aux catalogues (fiches, puis informatisations) ;
- le livre qu'empruntait un lecteur était le même que celui qui était emprunté ailleurs.

Le métier des bibliothèques a changé (2)

- aujourd'hui, les bibliothèques créent davantage : elles numérisent ;
- elles le font essentiellement pour trois raisons :
 - elles possèdent les livres ;
 - le matériel de numérisation est relativement simple et peu onéreux (appareil photo plus informatique)
 - il est facile de faire une numérisation acceptable
- si les numérisations étaient plus difficiles, elles seraient beaucoup plus centralisées

Le métier des bibliothèques a changé (3)

La numérisation s'est naturellement retrouvée dans les mains des bibliothèques :

- les bibliothèques font les choix
- l'utilisateur est presque inexistant de cette chaîne :
 - principalement parce que les numérisations ne correspondent pas à un besoin préexistant ;
 - les utilisateurs ne sont pas exigeants.
- les bibliothèques ont eu et ont encore carte blanche.
- mais ... cela conduit à des problèmes.

Une partie de ces problèmes provient d'un **éclatement des rôles**.

L'éclatement des rôles dans la numérisation

- ceux qui numérisent ne sont pas ceux qui utilisent ;
- ceux qui numérisent ne sont pas à l'écoute des utilisateurs ;
- ceux qui numérisent ne sont pas techniciens ;
- ceux qui décident ne numérisent pas ;
- ceux qui programment ne numérisent pas.

Ceux qui numérisent ne sont pas ceux qui utilisent

Conséquences prévisibles et visibles :

- pas de connaissance des véritables besoins
- pas de prise en compte de la qualité de numérisation
- pas de réelle compréhension des objectifs de qualité et de confort

Ceux qui numérisent ne sont pas à l'écoute des utilisateurs

- Traditionnellement, les bibliothèques ne consultaient pas/peu les utilisateurs pour faire des choix.
- Cette habitude s'est perpétuée.
- Mais elle n'est plus adaptée au monde numérique.

De plus, l'écoute des utilisateurs est faussée :

- pour certaines bibliothèques, l'absence de retours légitime les choix
- pour certaines bibliothèques, quelques retours positifs légitiment les choix

Ceux qui numérisent ne sont pas techniciens

- pas de formation sur la capacité limite de numérisation
- pas de vision à long terme
- pas de volonté d'en finir avec le chaos
- les numérisateurs n'ont pas de formation en physique, et ne sont pas des informaticiens
- les numérisateurs utilisent des appareils de numérisation, puis manipulent des outils de retouche et de gestion d'images

Ceux qui décident ne numérisent pas

- et ils n'utilisent pas non plus les livres numérisés
- par conséquent, ils ne voient pas certaines erreurs, ou ne réalisent pas que ce sont des erreurs

Ceux qui programment ne numérisent pas

- ils ne lisent pas ce qui est numérisé
- et ils sont trop bas dans la chaîne pour faire remonter d'éventuelles remarques
- lorsque des erreurs sont remarquées, il est trop tard, le livre est rangé, il est trop coûteux d'y revenir

Interactions entre les bibliothèques et Google

Les bibliothèques ont été abusées par Google :

- est-il possible que tant de bibliothèques prestigieuses aient accepté que Google emploie un si faible niveau de qualité ?
- les dirigeants des bibliothèques ignorent en fait ces problèmes !
- un conservateur de patrimoine sérieux se doit de refuser que les numérisations soient bâclées à ce point
- Google connaît les erreurs (et pourrait les corriger), mais réussit quand même à se faire accepter
- Google ne se gêne pas pour vendre des ouvrages incomplets
- pour beaucoup de bibliothèques, une qualité moyenne est compensée par une grande quantité à coût presque nul
- la BM de Lyon a commencé à être numérisée par Google en 2009, on peut s'attendre à la même catastrophe

Problèmes de recul

Les bibliothécaires :

- pour eux, les utilisateurs sont satisfaits (en fait, les utilisateurs n'utilisent pas)
- chaque technologie est remplacée par la suivante (pas de notion de durabilité)
- les moyens sont limités
- les sous-traitants ont dit ...
- ils se fient à des historiens (chartistes et autres) non techniciens
- ils ignorent les détails de Google (puisque bibliothécaires et utilisateurs sont différents) et s'en désintéressent même

Les fausses priorités des bibliothèques

- certaines bibliothèques jurent beaucoup trop par les méta-données
- la priorité première doit-être la fidélité du contenu, le titre, l'auteur, l'année
- on peut faire de l'OCR, mais dans ce cas-là il est souhaitable de le corriger (et l'OCR peut être externalisé)
- l'OCR ne doit pas être un préalable à la mise en ligne (ce que font certaines bibliothèques)
- il ne faut pas numériser pour simplement montrer qu'on a une activité de numérisation
- en se focalisant sur les méta-données, les bibliothèques passent à côté des problèmes de qualité de numérisation

Les fausses priorités des bibliothèques (2)

L'impression qui se dégage, c'est que les bibliothèques

- voient la numérisation comme une sous-traitance
- estiment que leur travail concerne l'enrobage (métadonnées, OCR, interface, etc.)

Cette conception crée une distanciation entre les bibliothèques et le livre.

Mauvaises numérisations et doublons

- un même ouvrage peut avoir été numérisé à plusieurs reprises ;
- Alain Jacquesson écrit (p. 75 de son livre) : « Nous ne pouvons que regretter que l'œuvre de Stendhal, Le Rouge et le Noir, ait été numérisée, dans une même édition, à la fois par l'université de Californie et par la bibliothèque de l'État de Bavière à Munich ».
- mais, faut-il vraiment le regretter ?
- en fait, **la question de Jacquesson est mal posée**, car elle ignore la qualité de numérisation ;

Les doublons ne sont un mal qu'à qualité équivalente ou inférieure.

Il faut d'abord comparer les qualités. Ne pas les comparer handicape les numérisations futures.

Mais ... les doublons sont quelquefois utiles

- Dans certains cas, des exemplaires en apparence identiques ne le sont pas.
- Un fonds hétéroclite (par exemple les archives de tel écrivain) peut contenir des documents déjà numérisés ailleurs, mais que l'on souhaite tout de même inclure dans le fonds, pour des raisons de cohérence.

Les doublons ne devraient pas être fortuits.

Le lourd héritage de la mauvaise numérisation

- les mauvaises numérisations bloquent la réalisation de meilleures numérisations
- à Strasbourg, par exemple, une numérisation qui a été faite ailleurs rend une nouvelle numérisation du même ouvrage moins prioritaire
- ce handicap est d'autant plus important que les responsables de la numérisation ne comprennent pas les exigences de qualité

Les problèmes de politique de numérisation

Buts de la numérisation

Au début, nous avons supposé que **l'objectif était la qualité**.
Cet objectif est toujours d'actualité.

Nous avons examiné les points suivants :

- quelle doit être la qualité de la numérisation ?
- comment les numérisations doivent-elles être présentées ?
- quelles sont les causes des problèmes rencontrés ?

Il reste maintenant à voir les choix des contenus :

que faut-il numériser ?

Comment sélectionner un fonds à numériser ?

Les numérisations se font habituellement par fonds. On peut envisager différents critères :

- fonds unique, par exemple des manuscrits ;
- fonds très demandé (essentiellement des manuscrits) ;
- etc.

Dans tous les cas, il faut

- avoir une connaissance des fonds avant de choisir ;
- avoir une politique à long terme, et non locale ;
- impliquer les utilisateurs.

Cahier des charges

Le cahier des charges que fait un organisme demandeur de numérisation sort du cadre de cette présentation, mais il faut qu'il précise :

- les conditions de maintien de l'intégrité du fonds (par exemple ordre et disposition des documents)
- les conditions techniques de la numérisation
- les conditions de stockage pérenne
- les conditions de mise à disposition des utilisateurs

Il est aussi souhaitable que ce cahier des charges ne soit pas confidentiel et soit librement accessible par les utilisateurs pour que ceux-ci puissent le commenter.

Stratégie pour un petit centre de numérisation

Un petit centre n'a pas les moyens des grands acteurs. Ses choix en sont affectés.

Une stratégie est alors de

- bien numériser ce que les autres ne feront pas bien ;
- bien numériser ce que les autres ne possèdent pas, par exemple des manuscrits.

Y a-t-il des choses à ne pas numériser ?

Cette question revient souvent, mais :

- c'est une mauvaise question ;
- tout ce qui a été conservé sous forme papier, parchemin, papyrus, etc., manuscrit ou imprimé, est en droit d'être numérisé ;
- la quantité de documents n'existant pas nativement sous forme numérique va certainement décroître dans les bibliothèques et archives, ce qui finalement limitera la quantité à numériser ;
- le problème de la non-numérisation doit être ramené en aval à celui de la non-conservation ; le numériseur n'a pas à se poser la question : faut-il ou ne faut-il pas numériser tel document ? s'il a été conservé, il doit être numérisé.

Le problème des choix

Les choix des fonds reposent souvent sur des malentendus :

- important historiquement \neq urgent à numériser ;
- il n'y a presque aucune urgence de numérisation ;
- les ouvrages numérisés ne sont pas lus.

Exemple de l'Université de Strasbourg

- le SCD de Strasbourg numérise des ouvrages depuis environ 2005
- le fonds scientifique est très riche
- les ouvrages sont choisis par des personnes de divers domaines, très souvent des professeurs retraités
- néanmoins, les choix semblent faits sans réelle connaissance des fonds ;
- en 2010, pour pallier ce problème, j'ai rassemblé en 250 pages un aperçu des fonds scientifiques de l'université ;
- cet aperçu devait initialement permettre de mieux choisir les ouvrages à numériser, mais aussi devait éviter de tomber dans la sous-traitance abusive
- toutefois, ce travail a mis en évidence les limites de l'approche thématique suivie par le SCD de Strasbourg ;

Les limites d'une approche thématique

En numérisant uniquement thématiquement et avec des moyens limités (100 à 200 livres par an) :

- un fonds riche ne pourra être qu'effleuré ;
- l'échantillonnage est subjectif et les choix ne correspondent pas forcément à des besoins ;
- l'approche thématique (quotas par disciplines) est satisfaisante pour les personnes intéressées de loin (on fait plaisir à tout le monde), mais au fond, on piétine partout ;
- en voulant privilégier la thématique, on néglige la nature et les besoins des documents ;
- une approche thématique sur des ouvrages existant ailleurs est souvent un gaspillage d'argent public.

La sous-traitance en numérisation

- certaines bibliothèques ont une activité de numérisation, mais en plus sous-traitent à d'autres ; c'est le cas de la BNF qui sous-traite certaines numérisations à la BNU de Strasbourg
- la sous-traitance est acceptable si elle ne nuit pas au fonds de la bibliothèque sous-traitante
- en 2009, le SCD de l'université de Strasbourg a tenté de servir de sous-traitant à une institution parisienne ; cette sous-traitance allait se faire aux dépens du fonds du SCD de Strasbourg et n'était donc pas acceptable ;
- ce genre de dérives résultent d'une absence de centralisation forte des numérisations.

Numérisations-vitrines

Les numérisations sont maintenant à la mode :

- budget limité \implies action limitée
- un fonds particulier est numérisé, à des fins de test
- la numérisation-vitrine est-elle durable ?
- le but de la numérisation-vitrine est-il de mettre en branle une action globale (nationale) ?
- il faut cesser de numériser pour montrer que l'on sait numériser, ou pour faire comme les autres
- il faut numériser intelligemment

La numérisation comme discrimination positive

C'est une stratégie de valorisation :

- il s'agit ici de mettre en valeur un fonds, en lui donnant une préférence de numérisation ;
- c'est donc ce que l'on peut appeler une **discrimination positive**.

Une telle numérisation encourage la recherche dans un fonds donné.

La durabilité en numérisation

- la numérisation doit produire un objet permanent
- le contenu peut avoir une durabilité très supérieure à celle de l'interface
- exemples de numérisations non durables :
 - toutes les numérisations qui n'ont pas été faites à partir des originaux
 - certaines numérisations de Gallica provenant de microfilms
 - numérisations de registres d'état civil dans des archives (au Bas-Rhin faits à partir de microfilms)
- lorsqu'il s'agit d'un livre, on peut éventuellement attendre qu'une autre bibliothèque le numérise mieux
- mais pour les manuscrits, il faudra que ce soit la bibliothèque qui en est propriétaire
- la durabilité sera atteinte lorsque la qualité de numérisation permet de reproduire celle de l'original
- les numérisations insuffisantes gaspillent de l'argent

La durabilité en numérisation

Avec la numérisation durable :

- la numérisation est confortable
- le recours à l'original n'est plus nécessaire

Les numérisations à la demande (1)

- en 2009 ou 2010, j'ai suggéré à une bibliothèque allemande de numériser un livre ancien qui ne l'avait pas encore été
- la bibliothèque m'a répondu que c'était possible et à quel coût
- question : qui décide des numérisations si un chercheur n'est pas écouté ?
- la bonne solution eût été
 - soit de soumettre cette suggestion à ceux qui font les choix, afin de voir si la suggestion est cohérente avec la politique de numérisation et de l'y intégrer ; dans le cas contraire, on peut éventuellement facturer l'utilisateur
 - soit de prévoir dans la politique un espace pour les numérisations à la demande, sans facturation
- j'ai finalement renoncé à la numérisation, préférant attendre qu'une bibliothèque la fasse

Les numérisations à la demande (2)

- en 2009 ou 2010, la BNU de Strasbourg a numérisé pour moi (et pour elle !) des ouvrages qui m'ont été facturés (et qui depuis sont disponibles sur d'autres sites...);
- cette facturation était alors naturelle, car la BNU ne semble pas avoir une politique de numérisation développée comme les bibliothèques allemandes

Exemple des mathématiques

- il existe une communauté qui s'intéresse à la numérisation des mathématiques
- on estime que le nombre total de pages du corpus mathématique est d'environ 100 millions
- projet « Digital Mathematics Library » (DML)
- la dernière conférence sur ce thème s'est tenue en juillet 2011.

Digital Mathematics Library (DML)

Thèmes de la conférence DML 2011 :

- o search, indexing and retrieval of mathematical documents
- o ranking of mathematical papers, similarity of mathematical documents
- o math OCR with MathML/TeX output
- o document conversions from/to MathML, OpenMath, ..., PDF
- o conversions between various mathematical formalisms
- o mathematical document compression, processing of scanned images
- o algorithms for crosslinking of bibliographical items, intext citations search
- o mathematical document classification, MSC 2010
- o mathematical text mining
- o mathematical documents metadata exchange
- o long term archiving, data migration
- o reports and experience from math digitization projects
- o math publishing with long term archival goal
- o software engineering aspects of creating, handling MathML, OMDoc, OpenMath documents, and displaying them in web browsers

Exemple des mathématiques

La liste précédente montre que

- ce sont essentiellement les mathématiciens qui s'intéressent à la numérisation des mathématiques
- l'objectif est de rendre accessible les mathématiques utiles qui ne seraient pas encore numérisées
- paradoxalement, les thèmes de la conférence évoquent très peu les documents anciens
- cette conférence ne représente en fait pas bien les intérêts de la communauté des historiens des mathématiques
- par conséquent, il est à craindre que la numérisation des mathématiques anciennes ne sera pas aussi bien défendue, car ces ouvrages sont à-peu-près inutiles de nos jours
- de tels problèmes apparaissent dans d'autres domaines, notamment en astronomie

Questions pour une numérisation

Plusieurs questions se posent en cas de numérisation :

- qui choisit le fonds à numériser ?
- pourquoi ce fonds ? y a-t-il une demande ?
- cette numérisation est-elle durable ?
- est-ce une numérisation-vitrine ?
- qui va numériser ? ces personnes comprennent-elles les limites de numérisation ?
- la numérisation va-t-elle être adaptée aux besoins des contenus (formules, images, schémas, etc.) ?
- qui choisit l'interface ? comment cette interface se situe-t-elle par rapport aux autres interfaces ?
- aura-t-on le téléchargement ? Si non, pourquoi pas ?
- aura-t-on le texte ? si oui, ce texte sera-t-il indexé sur le web ?
- le fonds numérisé pourra-t-il être reversé dans un fonds plus large de manière homogène ?

Solutions

Nous avons déjà vu des solutions précises à deux problèmes :

- problèmes de qualité de numérisation ;
- problèmes de qualité d'interfaces.

Il reste maintenant essentiellement à résoudre les problèmes politiques.

Le chaos de la numérisation

L'un des problèmes actuels est qu'il y a une certaine anarchie, à la fois dans ce qui est produit et dans ceux qui le produisent.

- surnumérisation (en quantité)
- différences de qualité, d'interface, de choix
- aucune prise en compte de la qualité
- problèmes de formation des responsables et de rôles

Qui doit choisir les documents à numériser ?

Pour éviter l'anarchie, il faut réglementer les numérisations :

- les bibliothèques ne devraient pas décider seules des numérisations
- les bibliothèques ne devraient avoir qu'un avis consultatif
- les spécialistes des domaines ne devraient pas non plus décider seuls des numérisations

Il faudrait en fait **planifier les numérisations globalement.**

Le chaos et ses solutions : mieux numériser

- idéalement, on devrait partout cesser de numériser et commencer à réfléchir
- notion de “couche” pour intégrer les multiples versions d'une numérisation (intégration de la qualité)
- prise en compte de la qualité : mal faire est acceptable, si on le prend en compte dans la chaîne de production
- définition d'une échelle de qualités
- cesser de faire des actions locales
- réfléchir à la durabilité de l'objet numérique, et privilégier la qualité et la réutilisabilité à la quantité
- comprendre les besoins des utilisateurs
- s'intéresser au livre

Mécanisme de sélection

Certains problèmes peuvent se résoudre tout seul par un simple mécanisme de sélection.

- l'évolution peut se faire par voie de concertation nationale ou internationale ;
- elle peut aussi se faire par simple sélection ;
- la multiplicité des numérisations va entraîner un mécanisme de sélection ;
- mais on n'aboutira pas naturellement au même résultat ; le libéralisme n'est pas la solution !
- le coût n'est certainement pas le même, puisqu'il y a beaucoup de pertes avec le chaos ;
- **les numérisations étant faites pour tous, avons-nous le droit de faire n'importe quoi ?**

L'utilisateur a changé

L'utilisateur d'aujourd'hui n'est plus le même qu'autrefois :

- autrefois, les livres d'une bibliothèque s'adressaient essentiellement aux lecteurs directs de cette bibliothèque (ville, université, etc.)
- aujourd'hui, les bibliothèques produisent pour un public national et même international

Les bibliothèques ont aujourd'hui des devoirs et une responsabilité beaucoup plus grande qu'avant.

Les frontières d'une numérisation ne sont plus celles d'une université, et pourtant certains responsables de numérisation ont encore ce schéma dans leur esprit !

Gallica, les bibliothèques allemandes, etc., ont tous des comptes à rendre à tous.

Différents types d'utilisateurs

- l'utilisateur typique, qui peut être exigeant, ne veut pas chercher un menu de configuration
- l'utilisateur typique est occasionnel ; il vient pour un seul ouvrage
- il est possible de satisfaire cet utilisateur moyen et beaucoup de sites sont assez satisfaisants ;
- mais certains, comme ceux basés sur Contentdm, ne le sont pas
- un utilisateur peut vouloir étudier un livre chez lui, sans connexion internet ; il faut qu'il puisse télécharger ou imprimer une version de bonne qualité.

Les bibliothèques frileuses ?

Comme nous l'avons dit plus haut, les bibliothèques donnent aujourd'hui l'image d'institutions vétustes :

- volonté de contrôler les documents en ligne (tatouage, stéganographie, etc.), afin de contrôler les usages
- restrictions au téléchargement
- attachement au modèle du prêt

La redéfinition du rôle des bibliothèques

- les bibliothèques ont un nouveau rôle à jouer : que feront les bibliothèques lorsque *tout* aura été numérisé ?
- elles conservent les originaux
- elles ne doivent pas craindre de donner les versions numériques
- le nouveau rôle des bibliothèques pourra être d'enrichir les documents (OCR, dossiers documentaires, etc.)

L'enrichissement des documents

- peut être séparé de la numérisation, à la fois temporellement et spatialement
- glissement vers la **pénétration des documents**
- c'est la valeur ajoutée fournie par les bibliothèques (comme les catalogues)

Parmi les formes d'enrichissement qui vont se développer :

- la saisie complète d'ouvrages avec mise en forme (projet Gutenberg et plus, projet LOCOMAT, etc.)
- la représentation structurée des ouvrages (XML, TEI, MathML, etc.)
- la réalisation de dossiers complémentaires pour ancrer un ouvrage sur la toile ; c'est un rôle de centre de documentation, mais qui va nécessairement pénétrer celui des bibliothèques.

Gestion de la qualité

La qualité des numérisations demande une gestion précise :

- chaque bibliothèque devrait maintenir une liste des numérisations à refaire
- la qualité de la numérisation peut être évaluée tout de suite, et elle peut aussi être évaluée en-dehors de la bibliothèque
- Gallica devrait par exemple prévoir une renumérisation de l'ouvrage vu précédemment (et bien d'autres)
- certaines bibliothèques le font peut-être, mais ce n'est pas visible

Une échelle de qualité simple (1)

- Nous proposons de donner une lettre de A à E pour la qualité de numérisation. On peut distinguer la qualité stockée et la qualité fournie, mais en général l'utilisateur ne peut juger que la qualité fournie.
- Cette échelle considère uniquement les images, et non la qualité du texte ou des métadonnées. C'est un autre problème.
- Il est illusoire de croire que tout sera transformé de si tôt par OCR, notamment les textes mathématiques et il y a donc un véritable besoin d'évaluation de la qualité des images.

Une échelle de qualité simple (2)

- A : la numérisation pourrait remplacer l'original (résolution supérieure ou égale à 600dpi, en couleur, bien alignée, planches bien numérisées, aucun défaut)
- B : correct, mais pas assez confortable, ou ne respecte pas les couleurs, ou autres petits défauts qui n'empêchent pas la lecture ; planches bien numérisées ; la recopie ne nécessite pas la compréhension ;
- C : un seul défaut grave (résolution insuffisante, planches non dépliées, déformations, etc.) ;
- D : deux défauts graves ;
- E : trois défauts graves ;
- etc.

Normalisation de la mise en ligne

- il est aujourd'hui temps de penser à une normalisation des mises en ligne ;
- des règles précises devraient être suivies par les grands numériseurs, tant pour la qualité des numérisations que pour la qualité des interfaces ;
- ceci est malheureusement encore loin d'être le cas, comme les exemples de ce document le prouvent ;
- par la suite, les normes pourront bien sûr évoluer.

Numérisation manuelle ou automatique ?

Il existe des appareils de numérisation tournant automatiquement les pages, mais :

- les documents anciens ne peuvent pas être traités sans risque ;
- Google n'utilise pas des appareils entièrement automatiques ;
- un traitement de qualité n'est actuellement pas envisageable de manière automatique, car **pour bien numériser, il faut savoir lire le contenu** (pour l'aligner, pour adapter la résolution, etc.) ;

Une partie des mauvaises numérisations est due au fait que ceux qui numérisent sont presque des machines.

Un bon numérisateur doit comprendre les besoins du livre.

La mort des numérisations

- le monde numérique va être peuplé d'un nombre croissant de numérisations et il y aura **beaucoup de numérisations multiples** (le même original ou le même livre plusieurs fois) dans des qualités différentes
- ce **foisonnement d'information** va nuire à la recherche, tout comme les « fermes de contenus » ont rendu aujourd'hui presque inutilisables les recherches sur internet
- les « fermes de contenus » conduisent les utilisateurs à se rabattre sur quelques sites connus, comme Wikipedia
- la même chose risque de se produire pour les numérisations, surtout si des **copies clandestines de bibliothèques** apparaissent, ce qui est à prévoir
- si la qualité est gérée, il faut alors **planifier l'élimination des mauvaises numérisations**

La problématique des microfilms

Il y a une différence fondamentale entre les microfilms et les numérisations actuelles :

- les microfilms n'ont jamais eu pour but de se substituer aux originaux, mais de les protéger ou de les rendre accessibles
- les microfilms sont presque toujours en N&B et de mauvaise qualité
- les numérisations actuelles sont en couleur et peuvent atteindre des résolutions rendant pratiquement inutiles l'accès aux originaux, sauf dans de rares cas

Cette différence d'objectif explique qu'il ait été historiquement acceptable de faire de mauvais microfilms.

Le problème du coût

- le coût conditionne les moyens
- les moyens pourraient conditionner la qualité
- en réalité, il ne faut pas faire de compromis sur la qualité
- les compromis sur la qualité résultent de faux problèmes d'urgence
- avec moins d'argent, il faut numériser moins
- il ne faut pas numériser moins bien

La physique de la numérisation et l'information utile

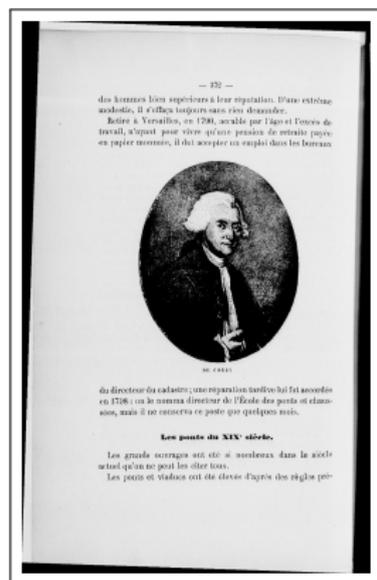
- notion de qualité limite atteignable et utile
- notions de base en traitement d'image (exemple : rotations)
- les caractères doivent être lisses, parce que nous sommes habitués à les voir lisses ; c'est notre **notion abstraite de caractères**
- les caractères lisses peuvent contribuer au confort
- pour le papier, **une résolution trop grande fait apparaître la fibre**
- l'information utile maximale est donc liée au support
- elle peut être limitée par les conditions de numérisation (optique, etc.)

Un mot sur les rotations d'images

- beaucoup d'images numérisées ne sont pas alignées ;
- il est possible d'appliquer une rotation aux images ;
- cette rotation entraîne presque toujours une perte de qualité (voir notre article sur le sujet) ;
- il est donc essentiel de minimiser cette étape de rotation, soit par un meilleur alignement, soit par une résolution beaucoup plus importante.

Exemple de rotation faite par Gallica

Debauxe : *Les travaux publics et les ingénieurs des ponts et chaussées depuis le XVII^e siècle*, 1893



- la page a été mal alignée à l'origine
- un traitement ultérieur l'a tournée, faisant apparaître les bords noirs en triangles
- cette rotation a fait chuter la résolution apparente de l'image

Formats de stockage : avec ou sans pertes ? (1)

- il y a actuellement souvent une fixation sur le format TIFF
- c'est un format qui peut conserver toutes les données de l'image initiale, mais il est volumineux
- dans une approche de « conservation », il paraît souvent évident aux bibliothèques d'utiliser ce format de manière interne
- mais, **c'est une fausse vision des choses !**
- si on choisit le format d'image en raisonnant de manière binaire (pertes ou non), on ne prend pas en compte la chaîne de numérisation
- le TIFF est une représentation sans pertes du produit d'une chaîne qui a perturbé l'objet (livre ou autre)

Formats de stockage : avec ou sans pertes ? (2)

- les sous-traitants pour la numérisation vont conseiller le TIFF, souvent par ignorance des autres paramètres
- certains utilisateurs veulent aussi du TIFF, parce que leurs éditeurs le leur demandent
- en demandant du TIFF, les éditeurs se prémunissent des erreurs dues à la compression, mais pas des erreurs en amont de la numérisation
- dans la presque totalité des cas, les bibliothèques n'ont en fait pas besoin d'utiliser ce format ; il ne cause pas de torts (à part en stockage), mais il n'est pas nécessaire
- **le vrai problème est de déterminer quelles pertes sont acceptables** vu les limites introduites par la chaîne de production (notamment limites optiques) et les limites physiologiques (résolution de l'œil)
- **l'absence de pertes est en fait une chimère**

Un mot sur Europeana

Europeana est essentiellement un **portail** vers les bibliothèques européennes :

- un peu l'analogue du SUDOC ou de KVK (Karlsruhe) pour les catalogues
- l'idée de faire un **catalogue unifié des collections numérisées en Europe** est excellente
- encore très incomplet (le SCD de Strasbourg, par exemple, n'y est pas)
- l'interface est assez simple
- pas de vision unifiée des œuvres
- les œuvres identiques ne sont pas rassemblées et leur qualité de numérisation n'est pas mesurée

Plusieurs gros problèmes

On peut distinguer actuellement plusieurs problèmes :

- il est souvent difficile de **savoir si un livre a été numérisé** quelque part ;
- **certains ouvrages sont numérisés en plusieurs exemplaires**, sans que ce soit toujours nécessaire ;
- **les numérisations sont éparpillées** sur de nombreux sites et **présentées de manière hétérogène**

Recensement des numérisations

Pour mieux maîtriser l'existence d'une numérisation, il est essentiel d'avoir un recensement global de tout ce qui a été numérisé.

- certains index existent, mais ils ne sont pas complets (exemple : <http://digreg.mathguide.de> à Goettingen)
- autre exemple : <http://www.zvdd.de/startseite>
- ces index s'appuient sur des listes fournies par les bibliothèques impliquées ;
- il n'y a certainement pas de notion d'un identifiant unique d'un exemplaire d'un ouvrage couvrant toutes les bibliothèques ainsi que les collections non cataloguées hors bibliothèques
- il est cependant probable que la réalisation d'un tel recensement est proche.

Liste de tâches de numérisation

Le recensement des numérisations n'empêche pas pour autant l'existence de numérisations multiples, puisque le recensement n'est pas forcément utilisé et que les doublons peuvent être simultanés. Pour résoudre ce problème, il faudrait avoir une liste de tâches de numérisation :

- identification unique des objets à numériser (livres ou parties de livres)
- affectation de tâches
- identification des ouvrages avec « marginalia » dont les annotations nécessitent une renumérisation, même si l'ouvrage lui-même a déjà été numérisé ailleurs

Serveur numérique central

L'un des avantages de Google est que l'interface est toujours la même, alors que les livres ne sont pas toujours au même endroit.

- les bibliothèques faisant de la numérisation devraient reverser toutes leurs numérisations à un serveur central
- la présentation de toutes les numérisations pourrait alors être unifiée
- Europeana n'est qu'un portail et ne fait que renvoyer aux différentes bibliothèques
- on peut penser que ce serveur numérique central verra un jour le jour
- Google a interdit à certaines de ses bibliothèques contractantes ce genre de reversement à partir des numérisations de Google, mais il n'est certainement pas interdit pour les nouvelles numérisations

La formation des responsables de la numérisation

- les responsables de la numérisation devraient davantage connaître les problèmes concrets
- ne pas se contenter d'une connaissance approximative fournie par des tiers
- pour être responsable de la numérisation, il faudrait par exemple être passé (au minimum) par l'exploitation et l'édition d'un ouvrage numérisé non trivial, qui force les candidats à utiliser les collections dont ils veulent être responsables
- de nos jours, certaines personnes deviennent responsables de la numérisation sans avoir fait aucun travail sur la numérisation ;
- faire des comparaisons d'interface, et d'autres études économiques, ne doit pas suffire ;
- le responsable de la numérisation doit être un visionnaire et ne pas travailler seul dans son coin ; la numérisation est un problème global.

Compétences requises

Un responsable de la numérisation devrait

- s'intéresser à tout le savoir
- avoir fait des recherches historiques au travers de livres et de manuscrits
- avoir des compétences techniques : en informatique, en physique, en optique (notamment physiologique), etc.
- comprendre les besoins des utilisateurs et interagir avec eux
- comprendre les besoins des livres
- s'impliquer dans des programmes globaux
- rechercher la qualité
- être visionnaire : avoir une vision à long terme, anticiper

Compétences requises (2)

- il n'est pas interdit d'avoir fait l'École des Chartes, ou l'ENSSIB, ou une thèse en histoire, ou encore de connaître les langues anciennes, mais cela joue un rôle très faible dans le travail du responsable de la numérisation
- par contre, si certaines compétences font défaut au responsable, il doit savoir s'entourer de personnes ayant ces compétences, et ne pas s'engager seul dans des actions
- certaines insuffisances actuelles viennent du fait que les profils de certains responsables ne correspondent pas aux compétences requises (problèmes de légitimité)

Et ... si on ne vise pas la qualité ?

Il est néanmoins possible de s'en sortir avec des numérisations moyennes, à condition :

- d'offrir la numérisation durable gratuite et rapide à la demande
- d'offrir la correction rapide des erreurs de numérisation.

Actuellement, aucune bibliothèque, et encore moins Google, n'offrent de tels service.

Faut-il pactiser avec Google ?

- un contrat avec Google permet d'avoir rapidement un grand nombre de numérisations de qualité moyenne ;
- il est possible d'améliorer cette situation en imposant à Google des clauses plus restrictives :
 - détection immédiate des erreurs (la presque totalité des erreurs peuvent être identifiées tout de suite) ;
 - soit on numérise un ouvrage intégralement (avec planches dépliées), soit pas du tout, mais pas partiellement ;
 - Google doit s'engager à ne pas mettre en ligne (et à ne pas vendre) de versions incomplètes ;
- dans tous les cas, même avec ces précautions, les bibliothèques contractantes doivent être conscientes qu'un jour (éventuellement lointain) la numérisation devra être refaite.

Perspectives

Aujourd'hui, **une grande majorité des numérisations effectuées dans les bibliothèques est à refaire**. On est loin de la durabilité !

- La qualité des numérisations croîtra lorsque les compétences seront mieux distribuées et que les besoins seront bien compris par tous.
- La quantité et l'ordre des numérisations croîtra lorsque la numérisation sera mieux organisée nationalement, au niveau européen et mondial, et pas seulement organisée par des versements aléatoires de numérisations non coordonnées.

Une note optimiste pour la fin ?

Le fait que les numérisations actuelles ne soient pas durables ne veut pas dire qu'elles ne doivent pas être faites :

- elles aident à faire prendre conscience des problèmes et à mieux définir les besoins (mais a-t-on besoin de n'avoir d'abord que des numérisations non durables ?) ;
- on pourrait faire le parallèle avec les générations de machines, les générations de portables (1G, 2G, 3G, etc.), etc., chaque génération étant utilisée pour accéder à la suivante ; mais est-ce vraiment la même chose avec les numérisations ?

Merci à toutes et tous pour vos échanges
et bonne numérisation !