

# HATHI TRUST RESEARCH CENTER

## **Text Mining HathiTrust as Big Data Exemplar in PRAGMA**

---

**Beth Plale – @bplale**  
**Professor, School of Informatics and Computing**  
**Director, Data To Insight Center**  
**Indiana University**



Tweet us - @HathiTrust #HTRC



# HathiTrust Digital Library

---

- HathiTrust is a partnership of academic & research institutions, offering a collection of millions of titles digitized from libraries around the world.
  - Founding members of HathiTrust along with University of Michigan are Indiana University, University of California, and University of Virginia



<http://www.hathitrust.org>

→ Distinguished  
from



<http://www.hathitrust.org/htrc>

#HTRC @HathiTrust



## Currently Digitized

- 10,796,403 total volumes
- 5,658,745 book titles
- 281,890 serial titles
- 3,778,741,050 pages
- 484 terabytes
- 128 miles
- 8,772 tons
- 3,450,939 volumes (~32% of total) in the public domain

View visualizations of HathiTrust  
[call numbers](#), [languages](#), and  
[dates](#)  
[statistics information »](#)

→ HathiTrust repository is a latent goldmine for text mining analysis, analysis of large-scale corpora through computational tools, and time-based analysis

→ Restricted nature of HT content suggests need for new forms of access that preserve intimate nature of research investigation while honoring restrictions

→ Paradigm: computation takes place close to the data



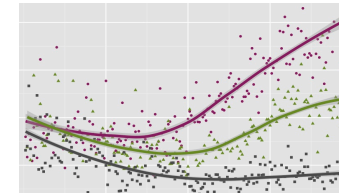
# Mission of HT Research Center

- Research arm of HathiTrust
- Goal: enable researchers world-wide to carry out computational investigation of HT repository through
  - Develop model for access: the ‘workset’
  - Develop tools that facilitate research by digital humanities and informatics communities
  - Develop secure cyberinfrastructure that allows computational investigation of entire copyrighted and public domain HathiTrust repository
- Established: July, 2011
- Collaborative effort of Indiana University and University of Illinois

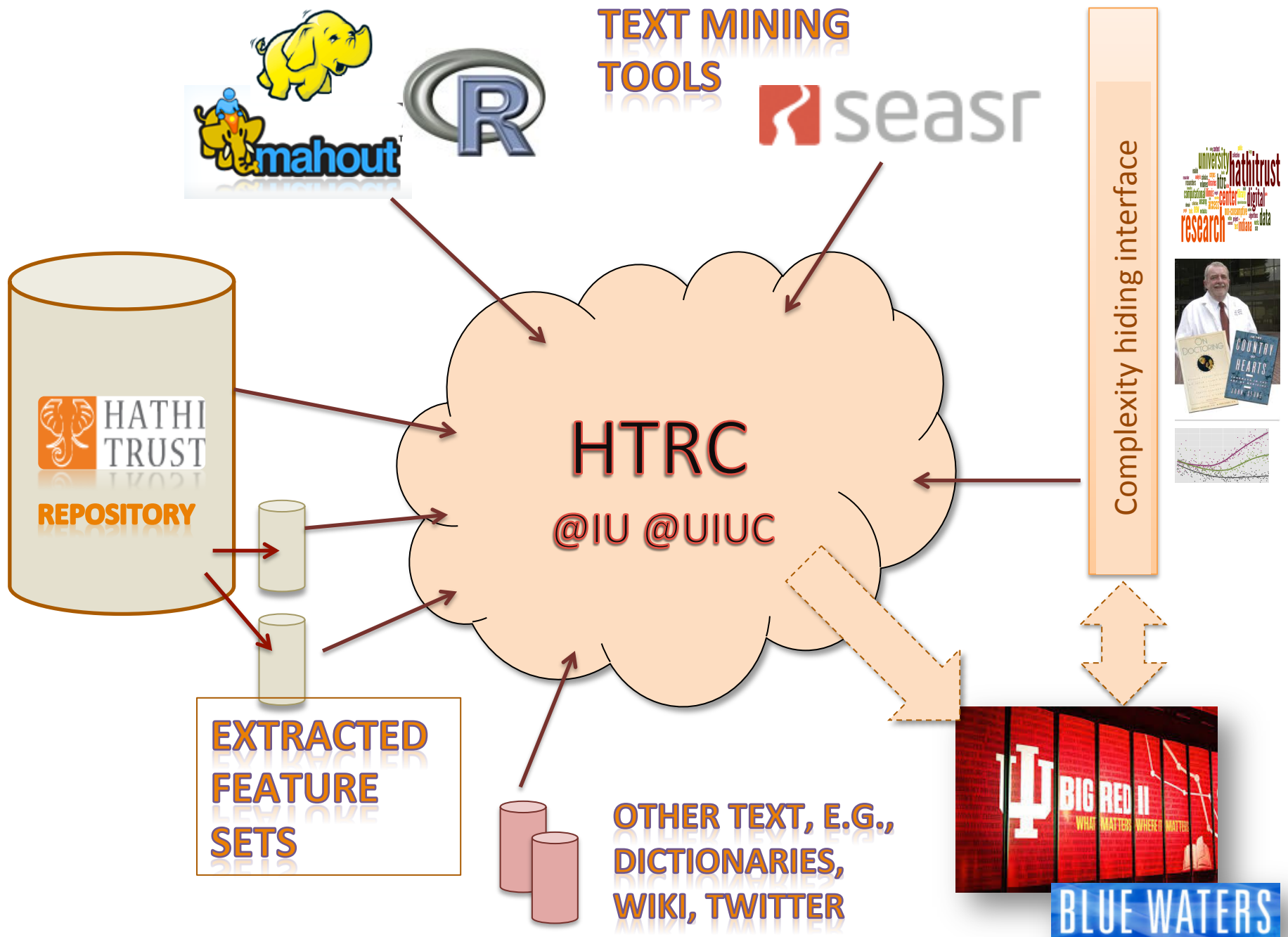


## Complexity hiding interface

## Tabular info



		2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035	2036	2037	2038	2039	2040	2041	2042	2043	2044	2045	2046	2047	2048	2049	2050	2051	2052	2053	2054	2055	2056	2057	2058	2059	2060	2061	2062	2063	2064	2065	2066	2067	2068	2069	2070	2071	2072	2073	2074	2075	2076	2077	2078	2079	2080	2081	2082	2083	2084	2085	2086	2087	2088	2089	2090	2091	2092	2093	2094	2095	2096	2097	2098	2099	2100	2101	2102	2103	2104	2105	2106	2107	2108	2109	2110	2111	2112	2113	2114	2115	2116	2117	2118	2119	2120	2121	2122	2123	2124	2125	2126	2127	2128	2129	2130	2131	2132	2133	2134	2135	2136	2137	2138	2139	2140	2141	2142	2143	2144	2145	2146	2147	2148	2149	2150	2151	2152	2153	2154	2155	2156	2157	2158	2159	2160	2161	2162	2163	2164	2165	2166	2167	2168	2169	2170	2171	2172	2173	2174	2175	2176	2177	2178	2179	2180	2181	2182	2183	2184	2185	2186	2187	2188	2189	2190	2191	2192	2193	2194	2195	2196	2197	2198	2199	2200	2201	2202	2203	2204	2205	2206	2207	2208	2209	2210	2211	2212	2213	2214	2215	2216	2217	2218	2219	2220	2221	2222	2223	2224	2225	2226	2227	2228	2229	2230	2231	2232	2233	2234	2235	2236	2237	2238	2239	2240	2241	2242	2243	2244	2245	2246	2247	2248	2249	2250	2251	2252	2253	2254	2255	2256	2257	2258	2259	2260	2261	2262	2263	2264	2265	2266	2267	2268	2269	2270	2271	2272	2273	2274	2275	2276	2277	2278	2279	2280	2281	2282	2283	2284	2285	2286	2287	2288	2289	2290	2291	2292	2293	2294	2295	2296	2297	2298	2299	2300	2301	2302	2303	2304	2305	2306	2307	2308	2309	2310	2311	2312	2313	2314	2315	2316	2317	2318	2319	2320	2321	2322	2323	2324	2325	2326	2327	2328	2329	2330	2331	2332	2333	2334	2335	2336	2337	2338	2339	2340	2341	2342	2343	2344	2345	2346	2347	2348	2349	2350	2351	2352	2353	2354	2355	2356	2357	2358	2359	2360	2361	2362	2363	2364	2365	2366	2367	2368	2369	2370	2371	2372	2373	2374	2375	2376	2377	2378	2379	2380	2381	2382	2383	2384	2385	2386	2387	2388	2389	2390	2391	2392	2393	2394	2395	2396	2397	2398	2399	2400	2401	2402	2403	2404	2405	2406	2407	2408	2409	2410	2411	2412	2413	2414	2415	2416	2417	2418	2419	2420	2421	2422	2423	2424	2425	2426	2427	2428	2429	2430	2431	2432	2433	2434	2435	2436	2437	2438	2439	2440	2441	2442	2443	2444	2445	2446	2447	2448	2449	2450	2451	2452	2453	2454	2455	2456	2457	2458	2459	2460	2461	2462	2463	2464	2465	2466	2467	2468	2469	2470	2471	2472	2473	2474	2475	2476	2477	2478	2479	2480	2481	2482	2483	2484	2485	2486	2487	2488	2489	2490	2491	2492	2493	2494	2495	2496	2497	2498	2499	2500	2501	2502	2503	2504	2505	2506	2507	2508	2509	2510	2511	2512	2513	2514	2515	2516	2517	2518	2519	2520	2521	2522	2523	2524	2525	2526	2527	2528	2529	2530	2531	2532	2533	2534	2535	2536	2537	2538	2539	2540	2541	2542	2543	2544	2545	2546	2547	2548	2549	2550	2551	2552	2553	2554	2555	2556	2557	2558	2559	2560	2561	2562	2563	2564	2565	2566	2567	2568	2569	2570	2571	2572	2573	2574	2575	2576	2577	2578	2579	2580	2581	2582	2583	2584	2585	2586	2587	2588	2589	2590	2591	2592	2593	2594	2595	2596	2597	2598	2599	2600	2601	2602	2603	2604	2605	2606	2607	2608	2609	2610	2611	2612	2613	2614	2615	2616	2617	2618	2619	2620	2621	2622	2623	2624	2625	2626	2627	2628	2629	2630	2631	2632	2633	2634	2635	2636	2637	2638	2639	2640	2641	2642	2643	2644	2645	2646	2647	2648	2649	2650	2651	2652	2653	2654	2655	2656	2657	2658	2659	2660	2661	2662	2663	2664	2665	2666	2667	2668	2669	2670	2671	2672	2673	2674	2675	2676	2677	2678	2679	2680	2681	2682	2683	2684	2685	2686	2687	2688	2689	2690	2691	2692	2693	2694	2695	2696	2697	2698	2699	2700	2701	2702	2703	2704	2705	2706	2707	2708	2709	2710	2711	2712	2713	2714	2715	2716	2717	2718	2719	2720	2721	2722	2723	2724	2725	2726	2727	2728	2729	2730	2731	2732	2733	2734	2735	2736	2737	2738	2739	2740	2741	2742	2743	2744	2745	2746	2747	2748	2749	2750	2751	2752	2753	2754	2755	2756	2757	2758	2759	2760	2761	2762	2763	2764	2765	2766	2767	2768	2769	2770	2771	2772	2773	2774	2775	2776	2777	2778	2779	2780	2781	2782	2783	2784	2785	2786	2787	2788	2789	2790	2791	2792	2793	2794	2795	2796	2797	2798	2799	2800	2801	2802	2803	2804	2805	2806	2807	2808	2809	2810	2811	2812	2813	2814	2815	2816	2817	2818	2819	2820	2821	2822	2823	2824	2825	2826	2827	2828	2829	2830	2831	2832	2833	2834	2835	2836	2837	2838	2839	2840	2841	2842	2843	2844	2845	2846	2847	2848	2849	2850	2851	2852	2853	2854	2855	2856	2857	2858	2859	2860	2861	2862	2863	2864	2865	2866	2867	2868	2869	2870	2871	2872	2873	2874	2875	2876	2877	2878	2879	2880	2881	2882	2883	2884	2885	2886	2887	2888	2889	2890	2891	2892	2893	2894	2895	2896	2897	2898	2899	2900	2901	2902	2903	2904	2905	2906	2907	2908	2909	2910	2911	2912	2913	2914	2915	2916	2917	2918	2919	2920	2921	2922	2923	2924	2925	2926	2927	2928	2929	2930	2931	2932	2933	2934	2935	2936	2937	2938	2939	2940	2941	2942	2943	2944	2945	2946	2947	2948	2949	2950	2951	2952	2953	2954	2955	2956	2957	2958	2959	2960	2961	2962	2963	2964	2965	2966	2967	2968	2969	2970	2971	2972	2973	2974	2975	2976	2977	2978	2979	2980	2981	2982	2983	2984	2985	2986	2987	2988	2989	2990	2991	2992	2993	2994	2995	2996	2997	2998	2999	3000
Utilities	Dong Hing Long Term Investment Securities, Inc. (DHL)	USD	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000	20																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																								

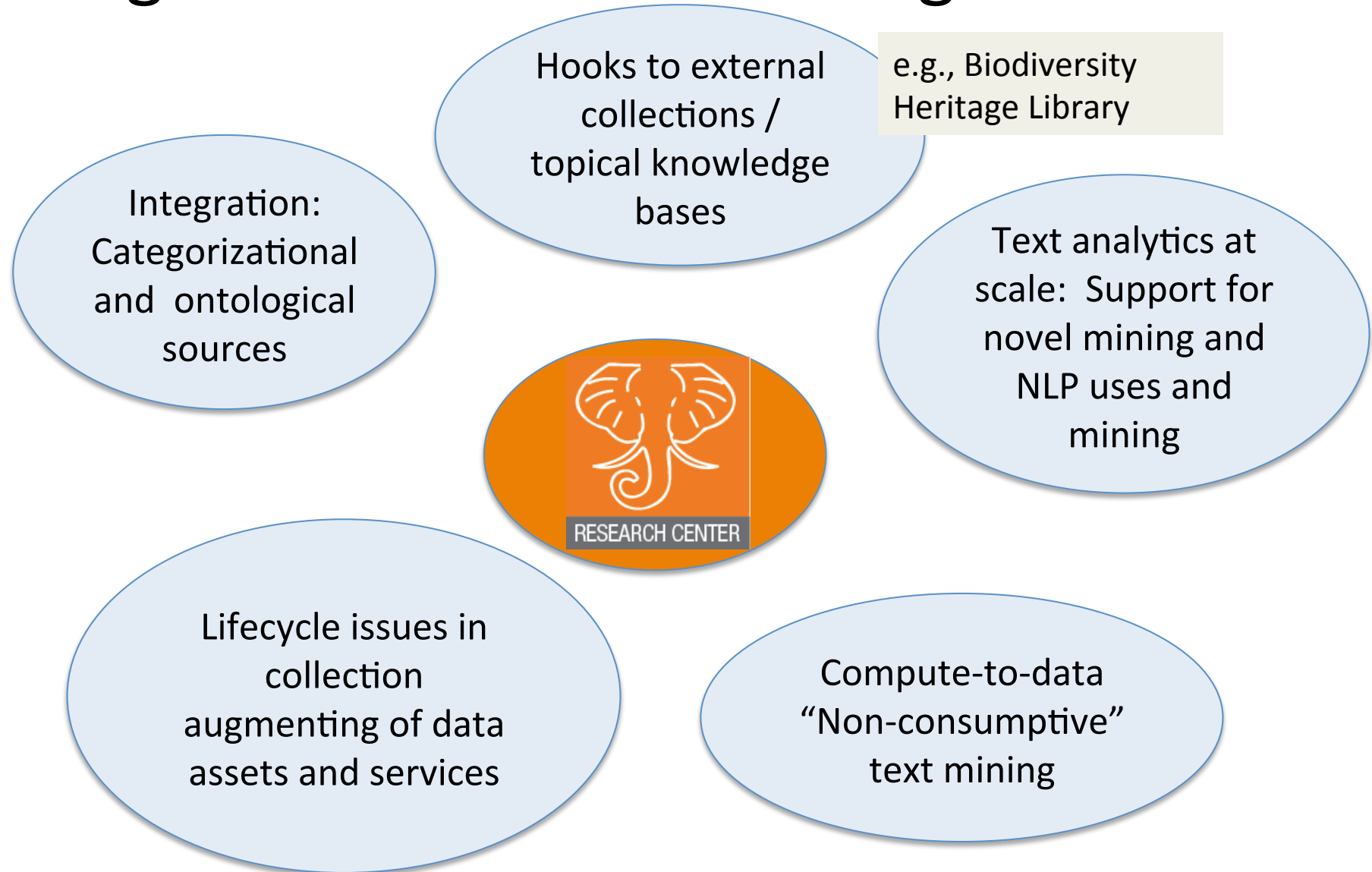


# HTRC architecture



- Philosophy: computation moves to data
- Web services (REST) architecture and protocols
- WS02 Registry for worksets and results
- Solr Indexes: full text, MARC, and new metadata
- noSQL (Cassandra) store as volume store
- Authentication using WS02 Identity Server
- Portal front-end, programmatic access
- Mining tools: currently SEASR

# Big Data Research Challenges in HTRC





# Big Data and PRAGMA

- Use PRAGMA member collections as anchor collections around which PRAGMA testbed could exercise experiments in data discovery and interoperability
- PRAGMA uses its testbed to exercise relevant activity of Research Data Alliance. RDA WG on Type Registries is about to deliver; assess usefulness in PRAGMA use cases
- Data on program for PRAGMA 27: devote time to address/expose policy and cultural issues in data sharing; are there tech solutions to policy issues that PRAGMA would be interested in?