Project HATCH

D7.1 UX Evaluation Results (Initial)

Project funded by the European Commission within Horizon 2020

Project no.:	776357
Project title:	HATCH
Instrument:	Horizon 2020
Thematic Priority:	Coordination and Support Activity (CSA)
Start date of project:	1 November 2017
Duration:	20 months
Leader:	Blue Dot Solutions (BDS)
Consortium:	WATERDOG mobile (WDM), WIT Berry (WIT), Catena Space (CSL), Eurecat (EUT)

D2.2 Detailed Architecture Document	
Document code	HATCH-D7.1(M9)
Summary	WP7/T7.5 – UX evaluation results: This document presents the rationale and methodology and results of the design validation tests
Dissemination level	PU
Delivery date	2018-08-31
Lead Beneficiary	EUT

Activity	**Role**	**Name**	**Date**
Prepared by	Task leader	Alex Pereda	2018-07-26
Verified by	WP leader	CSL	2018-07-31

Executive summary

This document presents the results of the UX evaluation activities carried so far in the project. This involved, the extraction of usability metrics the planning and implementation of the testing strategy, the pilots to crash tests the evaluation tools, and the actual user evaluation of the portals visual design.

This document is the first iteration of two of the deliverable, the second will include the usability tests and final conclusions. The results presented in what follows are very satisfactory in terms of the confidence they provide on the visual appeal of the portal's design, and provide some interesting research results and questions with regard to the use of physiological metrics in website UX evaluations.

List of changes

Version	Date	Change
0.1	2018-07-26	First draft version.
0.2	2018-08-29	Final draft version

Contents

1. Introduction

1.1 Purpose

This document is the first of two manifestations of the Deliverable 7.1, introduced on the HATCH project proposal. Its objective is to present the results of the **first user evaluation phase involving aesthetic evaluations of the homepage design.**

HATCH aims to be a user-friendly and visually appealing knowledge oriented web portal to become the main reference and entry point for European citizens and professionals interested in space research, therefore, evaluating how users perceive this aspects is key to the project development. In what follows we introduce the methodological approach to these evaluations as well as the results of the first phase of the evaluation, the validation of the aesthetic design aspects of the portal.

1.2 Scope

The use case definitions and user requirements are described in deliverables D1.1 and D1.2 respectively.

Based on the outcomes of WP1, EUT identified the use cases most suitable for using in the tests, as well as the most adequate general methods for UX validation. Departing from this the overall approach to testing in the project was scheduled as described in this document. Work in task 7.5 involved so far:

- The identification of relevant evaluation metrics, derived from the requirements and use case definitions,
- Definition of methodologies and general testing strategy
- Experimental design of the lab tasks and questionnaires
- The realization of the piloting tasks
- The main evaluation of perceived aesthetics of the homepage.

Dependencies:

- Use Case Definitions (Task 1.2)
- User Requirements Specification (Task 1.3)

Dependants:

- Frontend Development (Task 4.2)

2. UX Evaluation

2.1 Requirements and use case definitions

A follow-up of the use-case definition process was carried in order to derive human factor related metrics from them, as a preliminary step in the work to be developed in task 7.5. To this end, every requirement and use case implying actions that end users should be able to perform were identified as human factor requirements.

Generally agreed usability goals to define are: effectiveness: the degree of success with which users achieve their task goals; efficiency: the time it takes to complete tasks; and satisfaction, that is, user comfort and acceptability; (see ISO 9241, part 11 'Guidance on Usability'[1]). These are most easily derived from the evaluation of an existing system. Other more detailed usability issues provide more specific design objectives e.g. understandability, learnability, supportiveness, flexibility and attractiveness.

[1] International Organization for Standardization. (1998). ISO 9241—Ergonomic requirements for office work with visual display terminals (VDTs). Part 11: Guidance on usability. (available from www.iso.org/ iso/en/CatalogueListPage.CatalogueList). Geneva, Switzerland; ISO.

Specifically, we identified which requirements are related to human factors, focusing on those derived from use cases where the user is the primary actor and then assigned each of these to one of those general UX metrics, which does not imply that these metrics would only entail the listed requirements, but further evaluations metrics as well. For example, UR-01 to the "Human Factor" requirement group and to "Effectiveness" as main validation metric. All the requirements implying actions the user should be able to perform, were identified as human factors requirements, with all of them falling under the umbrella of effectiveness (success in completion), although some of those metrics can also involve efficiency (time to completion). The most relevant use cases were:

- UC-G-01 Search
- UC-G-02 Modify Search (P2)
- UC-G-03 Inspire
- UC-G-04 Share Shortcut
- UC-G-06 Access General Materials
- UC-G-08 Retrieve HATCH FAQ
- UC-U-05 Update User Profile
- UC-U-06 Manage Communication Preferences
- UC-U-09 Follow Object
- UC-U-10 Save Search
- UC-U-11 Manage Searches
- UC-U-12 Create Network
- UC-U-18 Reset Access

The use cases related to searching for research results seem the best option for effectiveness and efficiency tests, as both citizens and profs put this as main requirement (see figure 1), and the related Use cases are general and top priority, although we will consider others such as "Inspire" and "Create Network"

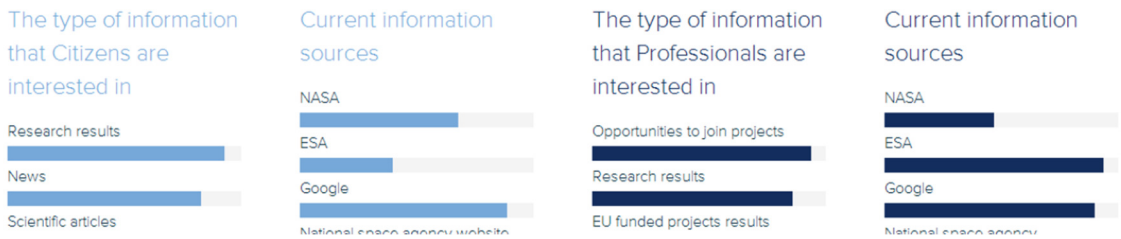

Figure 1. Answers to the requirements surveys regarding the key use cases.

2.2 General methodology

Broadly speaking there are two main kinds of evaluations to consider in this project:

- **Layout and design**: these involve mainly aesthetic and look&feel questions to test the perception of a given design, the tests mainly involve measures of aesthetic appeal and visual tests such as the first-impression test.
- **Effectiveness and efficiency**: these involve having participants executing the different actions identified as relevant in the use case definitions and requirements, while registering how well they do and how long they take, together with some self-reported tests of satisfaction related metrics), as well as measures of navigation, information complexity and interactivity.

A wide collection of methods is available to measure UX, grounded on different methodologies depending on the type of question explored. First, to analyse conscious processes, there are

standardised questionnaires for measuring perceptual aspects, perceived usability[2], cognitive working load[3] or affective reactions[45]Nevertheless, since it is unlikely that people can report information about processes over which they have little or no awareness[67], psychological research has traditionally favoured methods that allow exploring unconscious or automated psychological processes. Such methods provide online, moment by moment information, and are not dependent on subject biases such as, for example, social desirability bias. A broad classification of these methods could be: (i) behavioural, that is, measurement of psychophysical thresholds, reaction times, motion, and eye tracking, and (ii) physiological, which entail measuring physiological changes in users in response to psychological stimuli. Physiological signals on the other hand, allow psychological measurements while users are interacting with the assessed technology, and have been favoured in media research[8910111213], and we will be using them in addition to the traditional qualitative measurements througout all lab tests in the project. Specifically, **qualitative metrics** will include Items for perceived aesthetics as proposed in the seminal study by Lavie and Tractinsky[14], while items for perceived usability were extracted from a modification of the Post-study Usability Questionnaire[15]. On the other hand, **quantitative or physiological metrics** will include spatial eye-tracking and physiological valence (EMG) as aesthetic related metrics, as well as fixation analysis of eye-tracking and physiological measures of mental load as usability related metrics include Alternative measures we migh also consider include pupil dilation, eye-blinks, and mouse/keyboard tracking.

Our general strategy involves comparing Hatch which other related and well established websites while obtaining qualitative and quantitative measures of user's perceptions, based on small controlled experiments with quantitative and qualitative measures, plus larger online qualitative surveys. In principle layout and design tests are more important in the first phase of the project, whereas effectiveness and efficiency tests will become more relevant towards the end of the project, when the portal's functionalities begin to be ready. However, made an attempt to include a preliminar evaluaiton of usabilty based on clickable wireframes, but as we will explain, a pilot task demonstrated that its limited functionality made them unsuitable for this purpose. Therefore, and not considering the pilot studies, **the main results to be described in this first iteration of the present deliverable involve mainly layout or aesthetic evaluations of the home page.**

[2] J. R. Lewis. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. Intl. J. Hum.-Comput. Interact., 7(1):57{78, 1995.

[3] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. Proc. Human Factors and Ergonomics Society Annual Meeting, 50(9):904{908, 2006.

[4] E. R. Thompson. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). Journal of Cross-Cultural Psychology, 38(2):227{242, 2007.

[5] M. M. Bradley and P. J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry, 25(1):49 { 59, 1994.

[6] A. S. Babrow. Theory and method in research on audience motives. Journal of Broadcasting & Electronic Media, 32:471{487, 1988.

[7] R. E. Nisbett and T. D. Wilson. Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84:231{259, 1977.

[8] N. Ravaja. Contributions of psychophysiology to media research: Review and recommendations. Media Psychology, 6:193{235, 2004.

[9] M. Soleymani, M. Pantic, and T. Pun. Multimodal emotion recognition in response to videos. Affective Computing, IEEE Transactions on, 3(2):211{223, 2012.

[10] M. Barreda-~ Angeles, R. Pepion, E. Bosc, P. Le Callet, and A. Pereda-Ba~nos. Exploring the effects of 3d visual discomfort on viewers' emotions. In IEEE Int'l Conference on Image Processing, 753{757, 2014.

[11] S. Koelstra, C. Muhl, and I. Patras. EEG analysis for implicit tagging of video data. 3rd Int'l Conf. In A, effective Computing and Intelligent Interaction and Workshops, 1{6, 2009.

[12] M. J. Eugster, T. Ruotsalo, M. M. Spap_e, I. Kosunen, O. Barral, N. Ravaja, G. Jacucci, and S. Kaski. Predicting term-relevance from brain signals. In Proc. 37th Int'l ACM SIGIR Conf. Research #38; Development in Information Retrieval, 425{434, 2014.

[13] Y. Moshfeghi and J. M. Jose. An effective implicit relevance feedback technique using affective, physiological and behavioural features. In Proc. 36th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, 133{142, 2013.

[14] Lavie, T., & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International journal of human-computer studies*, *60*(3), 269-298.

[15] Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*, *14*(3-4), 463-488.

Finally it is worth mentioning that the results of al the tests to be carried in the project will allow us to adress two important research question in the field of UX metrics, namely, the relation between qualitative, eye-tracking and physiological valence with aesthetic metrics; and the relation between qualitative, eye-tracking and physiological arousal with usability metrics

2.3 Pilot tests

Prior to the realization of the main test, we carried two pilot tests in order to validate our pipelines for the collection, synchronization and analysis of data, as well as the selected qualitative factors. Pilot 1 focused on validating our choice of qualitative metrics, and involved eight naïve participants. Pilot 2, in addition to further validating qualitative metrics, was a crash test of the functionality of the clickable wireframes and the quantitative metrics, and involved a different group of eight naïve participants. The validation was positive in both pilots, participants understood the questions, results were in the expectable range, and the processing pipelines were shown to be ready for the main test. However, pilot 2 also demonstrated that the clickable wireframes were not functional enough for usability testing, and based Following is a description of the actual results of both pilots.

2.3.1 Pilot 1. Test 1

The first was a **"first impression" test** involving eight naïve participants whom inspected a blurred version of the Hatch home page, plus another four homepages selected from a public heuristic selection of best websites in terms of UX (worldbest.com), we also included space.com which was selected by us because of its similarities with Hatch. The logic behind these first impression tests is that first impressions and low level perceptual features are key in the overall evaluation of the aesthetics of any given stimulus. Thus, participants scored on a 5 point scale their general aesthetic impression based on coloring and layout. Figure 2 shows a detail of the blurred hatch page as an example.

Figure 2. Detail of the blurred version of Hatch's homepage.

Though the main idea of these tests is to validate our own tools, results showed that Hatch was well evaluated by the participants , it comes third, but it is worth noting that the competitor websites were selected solely on their being considered the highest level of aesthetic design. In the main test, as we will explain, we opted for a "fairer" set of competitor websites.

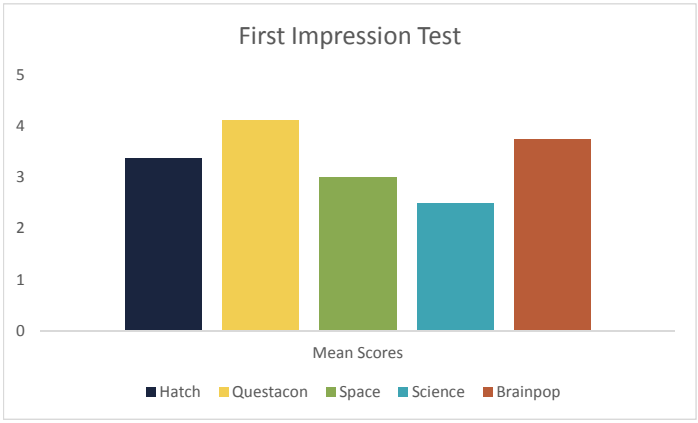

Figure 3. Results of the first impression test in pilot 1

2.3.2 Pilot 1. Test 2

The second test involved evaluating both versions of the Home Page in terms of the **main dimensions involved with aesthetic assessments**, as extracted from the literature referred in the introduction. Participants explored the page at will and scored the following factors in 5 point scales: Originality, Cleanness, Sophistication, Clarity, Fascination, Creativity, Pleasantness, Symmetry, and Attractiveness. As figure 4 shows, Hatch's main strengths seem to be **cleanness, clarity, pleasantness and attractiveness**. Importantly, results did not differ for both versions of the website (figure 5), so from this point on, we only used the main version.

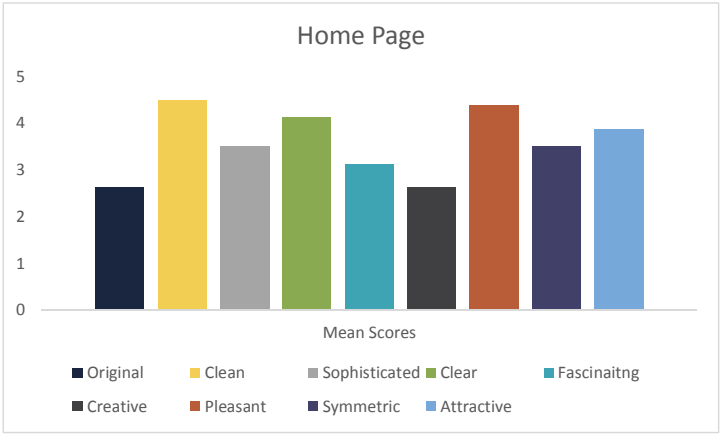

Figure 4. Aesthetics factors assessment in pilot 1.

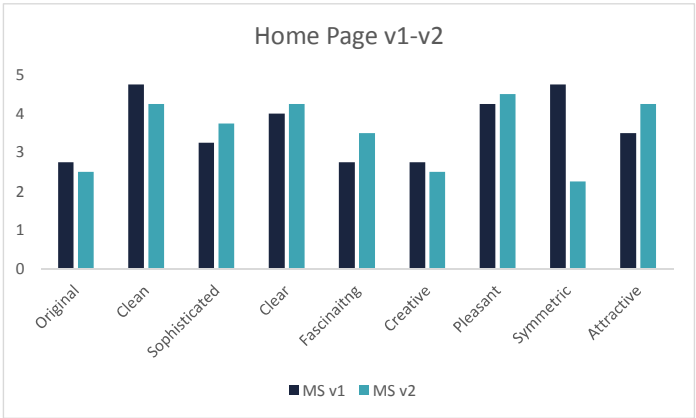

Figure 6. Aesthetics factors assessment in pilot 1 by version.

2.3.3 Pilot 2.

Pilot 2 involved testing of the clickable wireframes for the "Search" and "Save search" use cases, while also doing a crash test on the data collection, synchronization and analysis pipelines. Participants executed an equivalent search and save tasks in the Cordis website and in Hatch's wireframes in counterbalanced order, and then assessed our selected set of aesthetic dimensions. From the outset it became clear that the lack of functionality of the wireframes made the comparison unfair, which also explains why, in this case, the assessments of both sites were more similar than expected (figure 7). Based on these results, we decided to focus the main test on the aesthetic evaluation of the Home page. Regarding the validation of the quantitative metrics related processes, all of them proved to be ready for the main test.

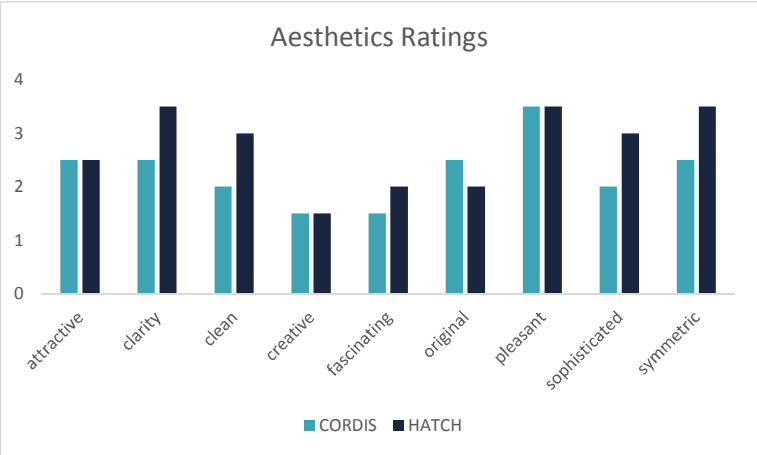

Figure 7. Rating of aesthetic dimensions in pilot 2

2.4 Main test

As explained above, the main test will address exclusively layout and aesthetic evaluations. It consisted of: consist of:

- Study 1: First impression test with blurred images of Hatch and the four benchmark sites gathering qualitative aesthetic metrics. **Cordis, NASA, ESA, and Space.com** were selected as final benchmark websites. Cordis was selected because it's the site that provides functionalities more akin to Hatch's, and will be key in the usability tests later on in the project. The other websites were selected because of their prominence and their thematic affinity with Hatch.

- Study 2: consists of a detailed evaluation of both Hatch and Cordis' homepages gathering qualitative and physiological measures.

Including both the online survey and the lab test, a total of 78 participants completed the survey (blurred + comparison cordis/hatch). They were 49 males and 25 females (and 4 non-declared), with ages between 17 and 68 (M = 32.27, SD=11.74). Regarding their highest academic level, 2 of them (~3% of the sample) had completed only basic studies, 29 of them (~37%) had a high school degree, 12 of them (~12%) had a college degree, 19 of them (~24%) had a master degree, and 12 of them (15%) had a Ph.D. Two participants did not declare their academic level, and two more select the option "other". They reported a range of occupations, being the most common the students (22, ~28% of the sample), IT and engineering related (13, ~16% of the sample), research and university teaching (9, ~12% of the sample), and business and managerial services (5, 6% of the sample), but also including other such as doctors, librarians, o blue collar workers. Regarding the lab study, the sample was composed of 22 participants (11 female) with ages between 21 and 50 (M = 27.59, SD = 8.29).

2.4.1 Main test – Study 1

As in the pilot, participants scored on a 5 point scale their general aesthetic impression based on coloring and layout of the blurred images. Order of presentation was counterbalanced. Results (Figure 8) show that both Hatch and Space.com are considered the most aesthetically pleasing, with no significant differences between them, but both of them being rated significantly higher than the other three.

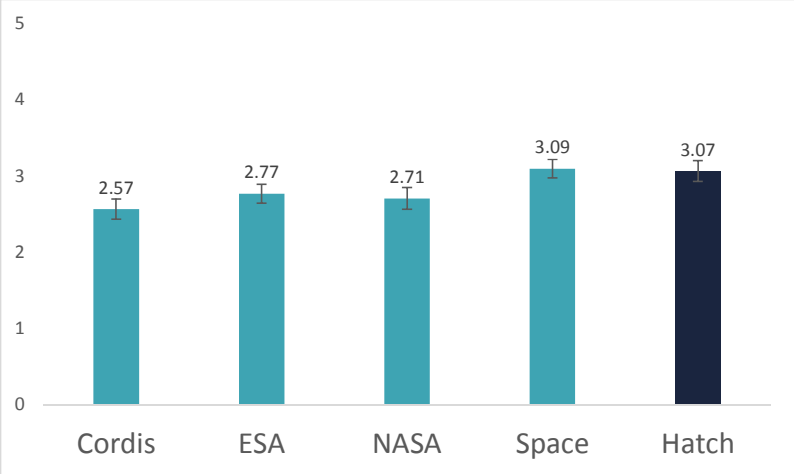

Figure 8. Results of the first impression test

2.4.2 Main test – Study 2

In this occasion, participants were presented with both homepages, in counterbalanced order, and told to explore them at will, without clicking on the mouse, for no less than one and no longer than three minutes, during which time eye tracking and EMG measures were collected. Then they were asked to rate each of the pages in terms of our selection of aesthetic dimensions, as well as a standard questionnaire for self-reported evaluations of perceived emotional state (self-assessment manikin). Results show that Hatch outscores Cordis in each of the assessed aesthetic dimensions, especially with regard to creativeness, fascination and originality, whereas the higher scores overall where obtained in clarity, cleanness and pleasantness, in line with the results of pilot 1.

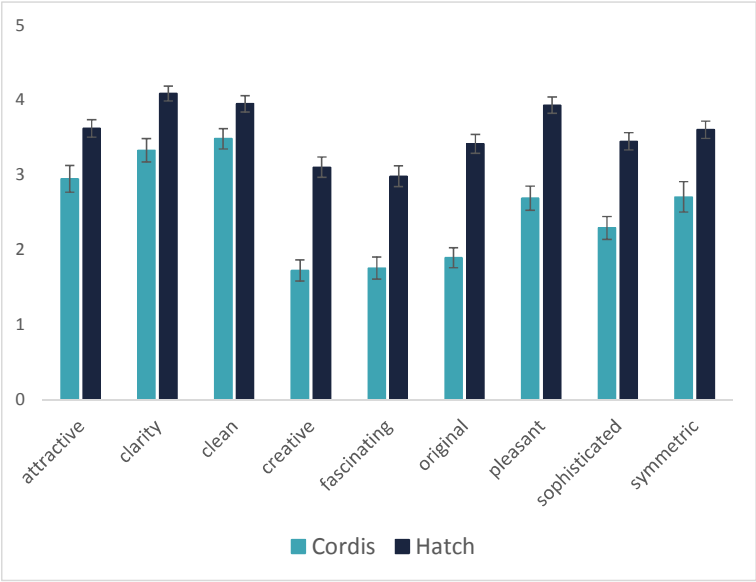

Figure 9. Rating of aesthetic dimensions in the main test

Regarding the self-reported evaluation of emotional state (figure 10) no differences were observed between sites, both were accompanied by a positive mood (valence) and a relaxed state (arousal).

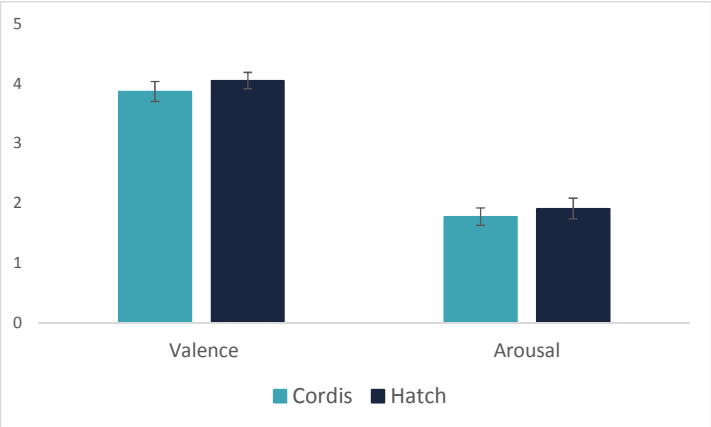

Figure 10. Self-reported valence and arousal during free exploration of the home pages

Turning now to the physiological measures (figure 11), EDA, the measure of activation and mental load (arousal) showed no variations between sites, which was expected given that free exploration in both sites was not supposed to draw differentially on mental resources. Contrary to our initial hypotheses however, EMG activation, our measure of pleasantness, did not vary between both sites although a non-significant tendency in the expected direction is observed, that is, responses of unpleasantness are higher for Cordis, it is worth noting that this also coincides with the self-reported emotional evaluation.

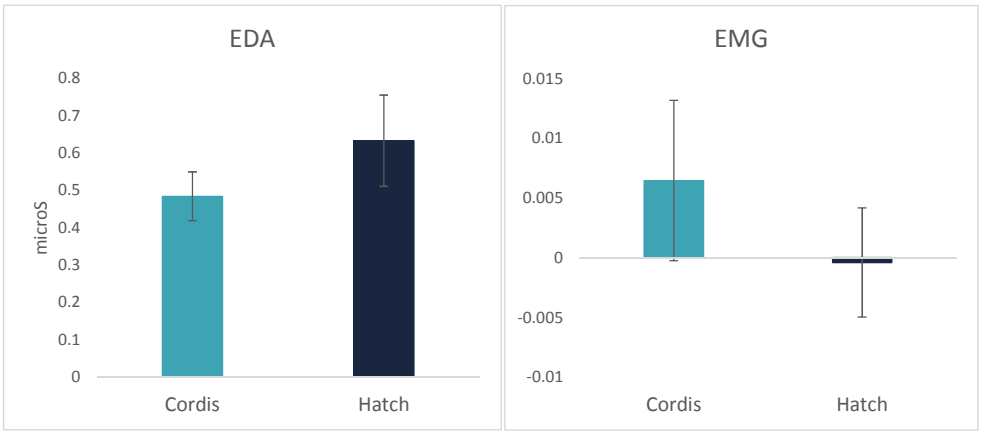

Figure 11. Physiological indicators of arousal (left) and valence (right)

Finally, regarding eye-tracking data, our main interest was to check whether there were variations in patterns of visual exploration tied to states of high and low valence. Given that there were no significant differences in this variable, the only conclusion we can extract for now from the data (figures 12 and 13) is that the pattern of exploration is similar for both sites and the expected in the case of a purely visual exploration in which participants do not have to process textual information or carry any actions. Both sites where explored evenly and no areas of the screen where ignored. In the next sections, we discuss the results altogether and present a final set of conclusions.

Figure 12. Visual exploration of Cordis' homepage

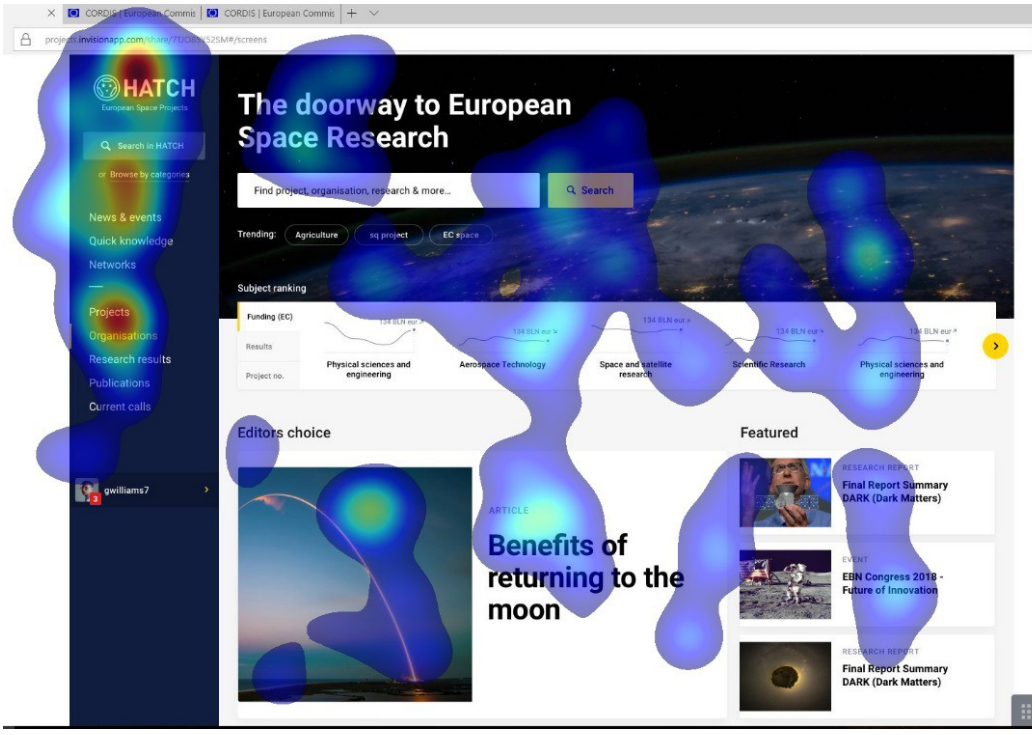

Figure 13. Visual exploration of Hatch's homepage

2.5 Discussion and conclusions

Results of the various studies here can be taken as evidence that **Hatch's visual design fully meets the desired aesthetic standards**, it fared well both when compared to both the selection in the pilots and main test. A recurring result is that **Hatch's design is perceived to be clear, clean and visually pleasing**, key aspects in making the site attractive especially for non-professional users. Low-level evaluations of

design and coloring, as evidenced by the various first impression tests also put Hatch at the level of top notch websites, especially when compared to those in with similar thematic content.

Regarding the physiological data, results were less clear cut, there is plenty of data in the literature reporting variations in self-reported valence correlating with the aesthetic appeal of websites, and ours was, to our knowledge, one of the first attempts to replicate these observations with physiological data, **thus the observation of no significant differences in EMG measures during aesthetic evaluations might come up us surprising**. It has to be noted however, that the decision of focusing the first phase of testing in layout/aesthetics aspects only might have played a role on the results, as when we designed the test, we expected that EMG measures will be obtained not with participants freely exploring the sites, which in the end is not a very natural activity, but during actual interaction. Thus, a putative hypothesis that we will be able to falsify with the data from the second phase, is that for effects of aesthetic appeal on EMG measures of valence, interaction with the website should be more natural. A hint to this was actually observed in pilot 2, the only result where Hatch did not score clearly higher than Cordis, precisely in an unfair comparison in which participants performed a search task, with Cordis being fully functional and Hatch only on clickable wireframes.

As for the eye-tracking data, even though we expected it to be more relevant in the usability tests, we were interested in observing whether differences in valence where related to a different pattern of visual exploration, but since no differences in valence were observed, this analysis was not granted. In any case, eye-tracking data confirms that **participants were indeed exploring both sites in a similar fashion and with the expectable patter for a non-interactive free visual exploration.** It is worth noting that despite the exploration was purely visual, the leftmost panel attracted most of the visual attention of the users, which validates and remarks the importance of this element in helping participants navigate the site. Again, we expect both physiological and eye-tracking data to acquire more relevance during usability testing, and if this in fact the case, the present results will contribute to clarify the role that these emotional processes play during both aesthetics and usability assessments.

To finalize, given that every aesthetic dimension was well scored **we would not recommend based on these results any significant modification of the visual identity of Hatch**. Even if we look at the lowest scored dimensions, creativeness and fascination, they were still above the mean and in informal conversations with the participants, they seem to be the least understood of the inquired dimensions, perhaps due to the fact that the interaction with the site was purely visual.

Next steps in the UX evaluation of Hatch involve preparing and scheduling the usability test, where we also expect to put to the test some of the research hypothesis that emerged from the present results.